# Assignment 2 – Data cleaning, preparation and transformation on Titanic dataset

*The purpose of this assignment is practice data preparation in Python*

This assignment provides you with an opportunity to demonstrate the achievement of the following course learning outcomes:

* Understand key data preparation steps
* Be able to manipulate data in Python
* Know most essential Python libraries and functions

## Key Information

* **Type:** *Individual*
* **Weight:** 6.25 %
* **Delivery:** Course website upload
* **Due Date:** End of lab session

## Expectations

You are expected to complete this assignment individually.

Respect for academic integrity is crucial to your success. Make sure you understand what constitutes acts of academic dishonesty in the page: [What is Academic Dishonesty?](http://mcmaster.ca/academicintegrity/students/whatis.html)

## Instructions

1. *Download Titanic dataset from Avenue to Learn*
2. *Manually delete values from several cells, save file locally and explore in Python*
3. *Apply following data preparation rules, note that additional documentation on Python is available at* [*https://docs.python.org*](https://docs.python.org)*,* [*https://matplotlib.org/*](https://matplotlib.org/) *,*[**https://pandas.pydata.org/pandas-docs/stable/**](https://pandas.pydata.org/pandas-docs/stable/)

*# Drop the columns where all elements are missing values:*

*df.dropna(how='all’)*

*# Drop the columns where any of the elements are missing values*

*df.dropna(how='any')*

*# Keep only the rows which contain 2 missing values maximum*

*df.dropna(thresh=2)*

*# Fill all missing values with the mean of the particular column*

*df.fillna(df.mean())*

*# Fill any missing value in column 'A' with the column median*

*df['A'].fillna(df['A'].median())*

1. *Look at the statistical summary of the prepared dataset:*

*Use: print( df.describe() )*

1. *Look at the statistical summary for categorical variables:*

*Use: categorical = df.dtypes[df.dtypes == "object"].index*

*print(categorical)*

*df[categorical].describe()*

1. *Identify categorical variables and explain why they cant be used in the raw form in the analysis*
2. *Transform one of the categorical variables into numerical using label encoder method*
3. *Transform gender into numerical variable using dummy coding*