

**ANL252**

**Python for Data Analytics**

# **End-of-Course Assessment**

**July 2021 Presentation**

**Submitted by:**

|  |  |
| --- | --- |
| **Name** | **PI No.** |
| **Peh Hong Wei Dyan** | **H1870791** |

**Tutorial Group: ­­­­­­­­­­­­ T09**

**Instructor’s Name: Mr. MUNISH KUMAR**

**Submission Date: 12/09/2021**

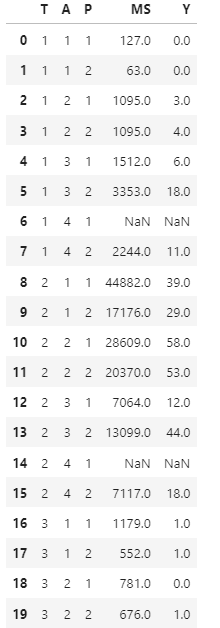
**1ai)**

!pip install pandas

import pandas as pd

ship = pd.read\_csv("ship.csv",na\_values = ".", na\_filter = True)

ship



**1aii)**

ship.columns = ["types", "c\_years", "o\_periods", "s\_months", "incidents"]

ship



**1aiii)**

import numpy as np

for i in range(1,6):

for j in range(1,3):

df = ship[(ship['types'] == i) & (ship['o\_periods'] == j)]

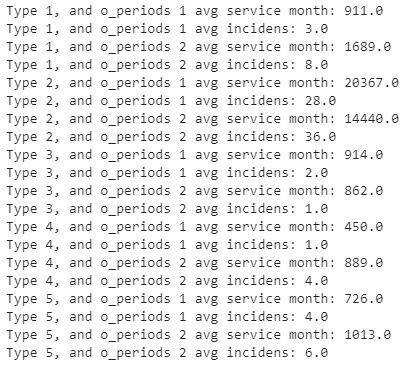
RSMth = float(df['s\_months'].mean())

Round\_RSMth = np.round(RSMth, decimals = 0)

RIncident = float(df['incidents'].mean())

Round\_RIncident = np.round(RIncident, decimals = 0)

print('Type ' + str(i) + ", and o\_periods " + str(j)+ ' avg service month: ' + str(Round\_RSMth))

print('Type ' + str(i) + ", and o\_periods " + str(j)+ ' avg incidens: ' + str(Round\_RIncident)) 

**1aiv)**

for i in range(1,6):

for j in range(1,3):

df = ship[(ship['types'] == i) & (ship['o\_periods'] == j)]

RSMth = df['s\_months'].mean()

RIncident = df['incidents'].mean()

ship['s\_months'].fillna(RSMth, inplace = True)

ship['incidents'].fillna(RIncident, inplace = True)

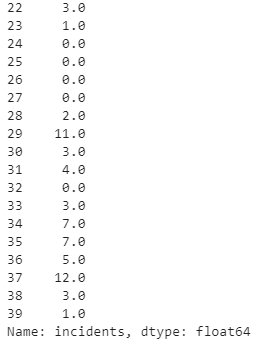
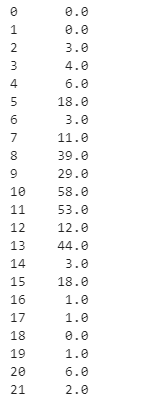
ship



**1av)**

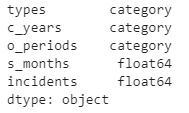
Y = ship['incidents']

Y



**1bi)**

df = ship.astype({"types":'category', "c\_years":'category', "o\_periods":'category'})

df.dtypes  


**1bii)**

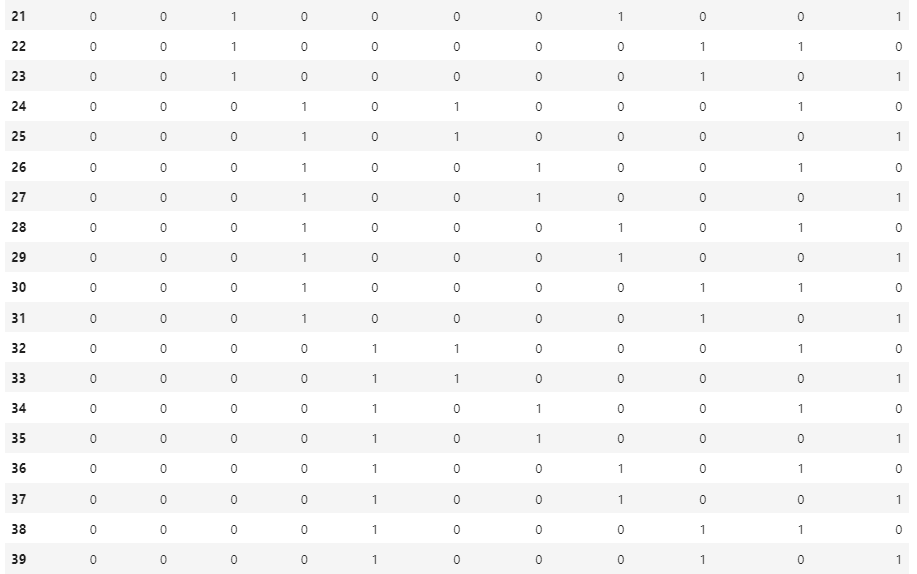
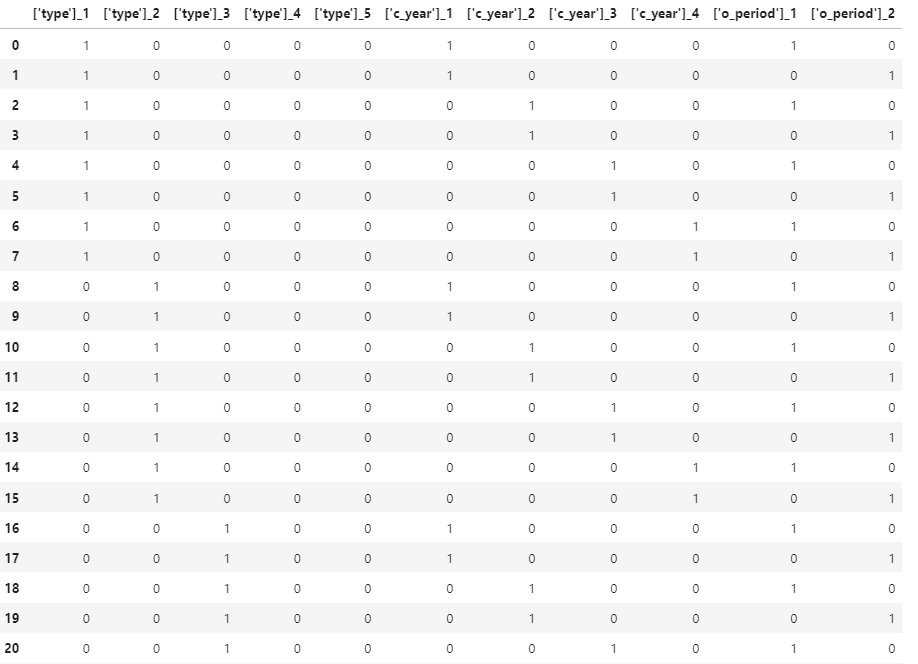
x1 = pd.get\_dummies(df["types"], prefix=['type'])

x2 = pd.get\_dummies(df["c\_years"], prefix=['c\_year'])

x3 = pd.get\_dummies(df["o\_periods"], prefix=['o\_period'])

X = pd.concat([x1,x2,x3],axis=1)

X



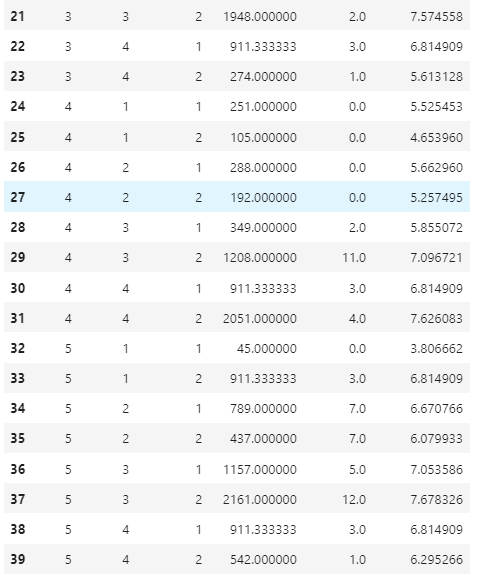
**1biii)**

import csv

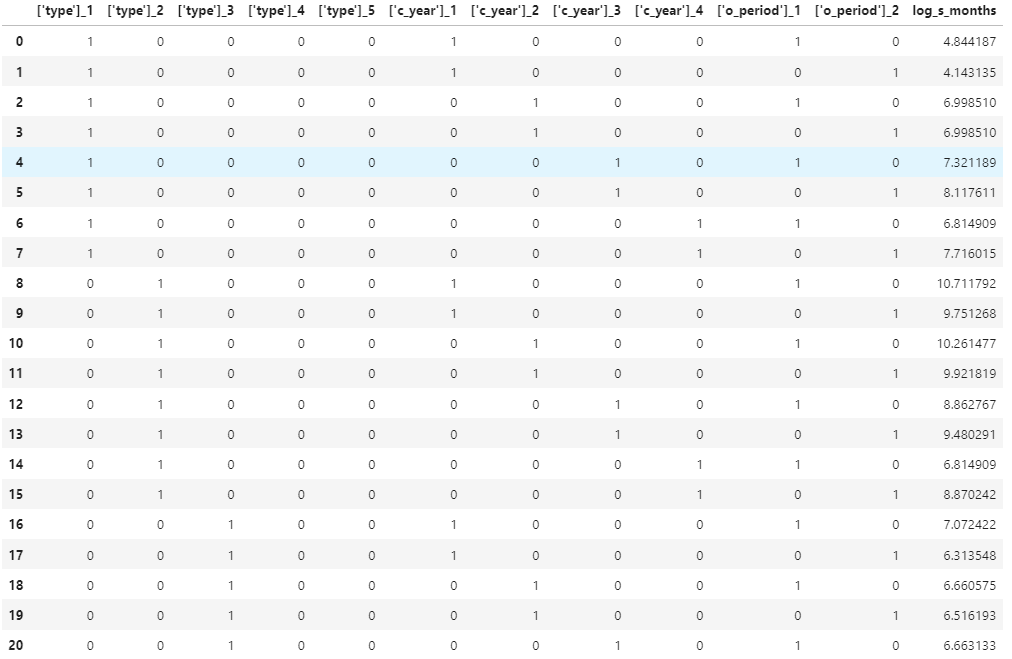
ship = pd.concat([ship,log\_s\_months],axis=1)

ship.columns = ["types", "c\_years", "o\_periods", "s\_months", "incidents", "log\_s\_months"]

ship



X = pd.concat([X,log\_s\_months],axis=1)

X.rename(columns={"s\_months" :"log\_s\_months"})  


**1c)**

The reason for not splitting the Dataframe into training and testing datasets to evaluate the predictive power of the model is the size of the dataset in the Dataframe.

Usually, when there is a large data in the dataframe, it will be better to split data into training and testing. Testing is a part of the data in the datasets use to test out and predict the model of the dataframe, as the bigger the dataset there will be more disruption to the finding and the model made. So, based on the data user could set a suitable amount of data to do the testing, instead of using all the data in the dataframe. As for the training dataset, it will be data that will be used to fit into the predicted model made using the testing data, to test the accuracy of the model made. (Lhessani, S., 2020)

However, the size of the ship dataset is only up to 40. So, it is relatively small, therefore there isn’t a need to split the dataset to predict the model.

**1d)**

with open('ship\_prepared.csv', 'w') as f:

fieldnames = ["types", "c\_years", "o\_periods", "s\_months", "incidents", "log\_s\_months"]

writer = csv.DictWriter(f, fieldnames)

writer.writeheader()

for i in range(40):

writer.writerow({"types": ship["types"].loc[i],"c\_years": ship["c\_years"].loc[i],

"o\_periods": ship["o\_periods"].loc[i],"s\_months": ship["s\_months"].loc[i],

"incidents": ship["incidents"].loc[i],"log\_s\_months": ship["log\_s\_months"].loc[i]})

import sqlite3

connection\_object = sqlite3.connect("ship\_db")

data\_object = pd.read\_csv("ship\_prepared.csv")

data\_object.to\_sql("ship\_data", connection\_object ,if\_exists = 'fail')

**2a)**

In scikit-learn the way to implement Poisson regression based on the question. First, we need to import sklearn, as it is the library that stores all the tools for data mining. Follow by data mining tool linear\_model from it, where it provides us to the different model for data mining. Using this tool we will be able to ask the system to use the Poisson regression module (e.g. linear\_model.PoissonRegressor()).

Next, to run the Poisson regression module, there are some of the parameters that we could add in the module base on the criteria (e.g. sklearn.linear\_model.PoissonRegressor(\*, alpha=1.0, fit\_intercept=True, max\_iter=100, tol=0.0001, warm\_start=False, verbose=0). Alpha represents regularization strength, which by default will be 1 if the parameters are not filled. Fit\_intercept is a Boolean test on the intercept. Max\_iter which is to set a maximum to the iteration of the model. Tol is criteria where the interaction will stop at once met. Warm\_start is to reuse the solution to fit as initialization for the coefficient and intercept. And, verbos is to set it to any positive number for verbosity. (Sklearn.linear\_model.PoissonRegressor, n.d.).

Lastly, is the fit() function, which will fit the training data and target value into the model. And, the predict() function which is used on an array of ample data, on the model made using the data in fit() and parameters set for the Poisson regression. (Sklearn.linear\_model.PoissonRegressor, n.d.).

**2b)**

!pip install sklearn-pandas

import sklearn

from sklearn import linear\_model

ship\_clf = linear\_model.PoissonRegressor()

ship\_clf.fit(X, Y)

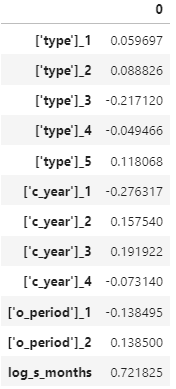
ship\_coef = ship\_clf.coef\_

lable = ["['type']\_1","['type']\_2","['type']\_3","['type']\_4","['type']\_5","['c\_year']\_1","['c\_year']\_2",

"['c\_year']\_3","['c\_year']\_4","['o\_period']\_1","['o\_period']\_2", "log\_s\_months"]

coefTable = pd.DataFrame(ship\_coef, lable)

coefTable



**2c)**

from sklearn.metrics import mean\_poisson\_deviance

y\_pred = ship\_clf.predict(X))

mean\_poisson\_deviance(Y, y\_pred)\*ship.shape[0]



Reference:

1) Wu, K. Y. (2021). *ANL252 Python for data analytics (study guide)*. Singapore University of

Social Sciences.

2) Lhessani, S. (2020, September 18). What is the difference between training and test dataset? Retrieved from <https://lhessani-sajid.medium.com/what-is-the-difference-between-training-and-test-dataset-91308080a4e8>

3) Sklearn.linear\_model.PoissonRegressor. (n.d.). Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.PoissonRegressor.html

4) Sklearn.metrics.mean\_poisson\_deviance. (n.d.). Retrieved from <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_poisson_deviance.html>