

**ANL252**

**Python for Data Analytics**

# **End of Course Assignment**

**July 2021 Presentation**

**Submitted by:**

|  |  |
| --- | --- |
| **Name** | **PI No.** |
| **Muhammad Farhan Bin Saad** | **H1771425** |

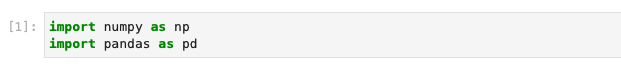
**Tutorial Group: ­­­­­­­­­­­­ T 09**

**Instructor’s Name: Mr. Munish Kumar**

**Submission Date: 13/09/2021**

**Question 1a.**

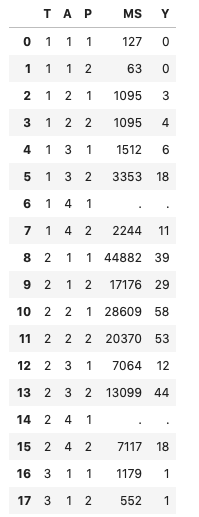
Before we begin building the program for this question, we need to import NumPy and pandas into the script as shown below in order to fulfil the following requirements. In addition, we need to upload ship.csv file into the Jupyterlab platform in order to conduct our assignment.

****

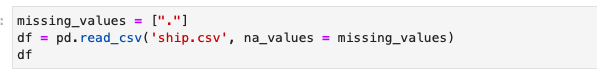
**Part i.**

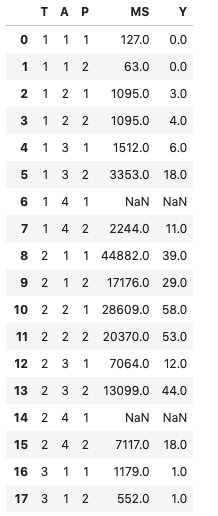
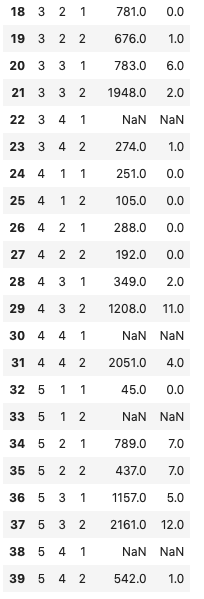
After uploading ship.csv file into Jupyterlab, we began reading the file using pandas shown below;

****

**** 

As seen from the data above, there 6 missing values which are shown as “.” and replace the values with NaN a term for missing values for Pandas, hence we are required to identify in Pandas (Yildirim, 2020) as missing values;



**Part ii.**

In this section, we are required to change the variable names as shown in the table;

|  |  |
| --- | --- |
| **T** | **Types** |
| **A** | **C\_Years** |
| **P** | **O\_Years** |
| **MS** | **S\_months** |
| **Y** | **Incidents** |

Using the data frame df.rename we insert the variables needed to change with the new variables shown below;



Hence obtaining the required variables shown below;



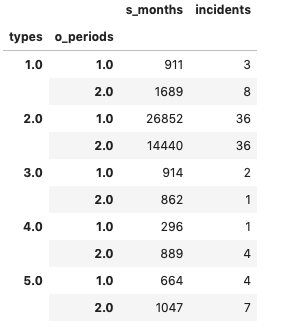


**Part iii.**

In this part, we would want to find the average of the service months(S\_months) the average number of incidents(Incidents) of every category of the types(T) and operation periods(O\_years) hence we input the following as shipgroup;

Taking note that we need to round up to the nearest integer



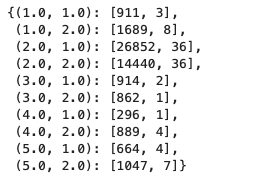


The corresponding output of the codes shown above shows the mean of S\_months and incidents corresponding to the category of the types and O\_periods respectively.

**Part iv.**

In this part of the question, we shall split into 2 parts. Firstly, we need to identify the values corresponding from the variable of ‘S\_months’ and ‘incident’ to the ‘types’ and ‘O\_periods’ which can be found in part (iii) and print as group\_dictionary in an array shown below;





Secondly, we have to replace the NaN value with the values taken from the array. We use the data taken from part (ii) and input an ELIF statement to compute the statement shown below;



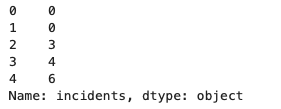




**Part V.**

As we only want the incidents column from the dataset, so we use loc (meaning location) to index the entire incidents column

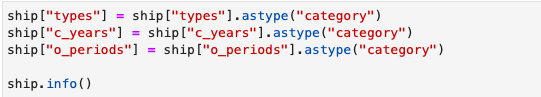


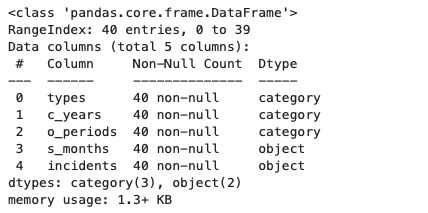


**Question 1b.**

**Part i.**

In this part, we need to input the dataframe.info (pandas, 2021) to identify categorically the information about the data frame of our case study.

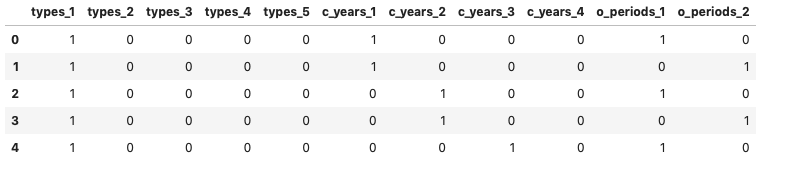
****

****

**Part ii.**

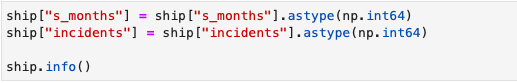
In this part, we input pd.get\_dummies (pandas, 2021) for the selected variables;

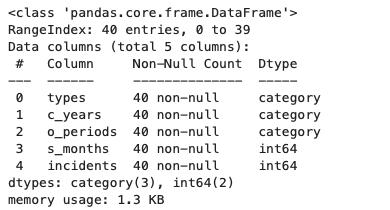
****

****

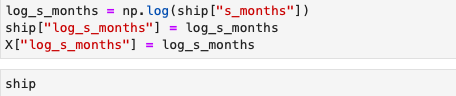
**Part iii.**

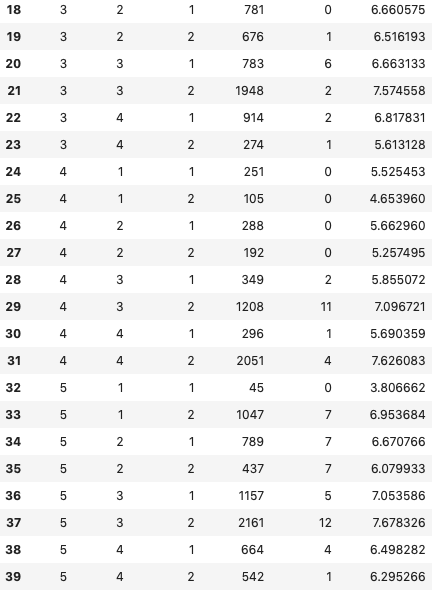
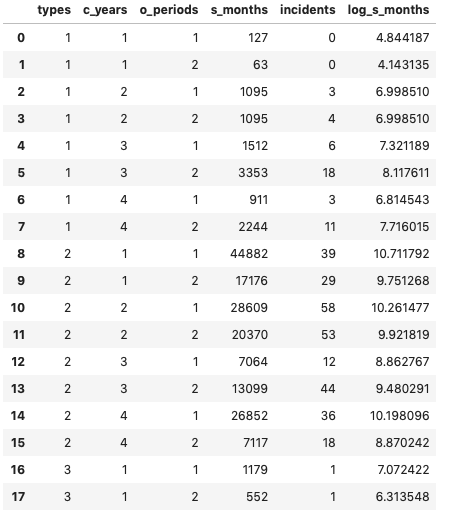
Similar to part (i) and part (ii), we will input dataframe.info for ‘S\_months’ and ‘incidents’ shown below;

****

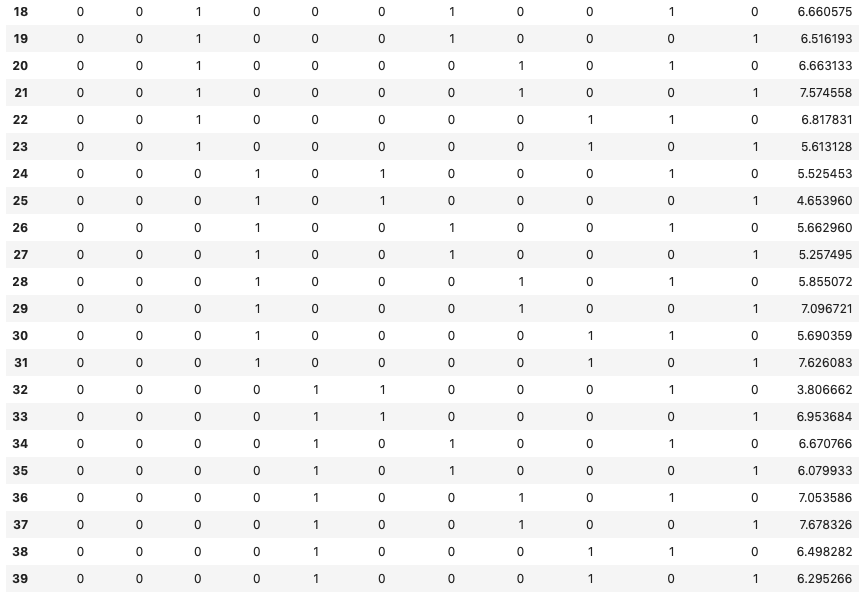
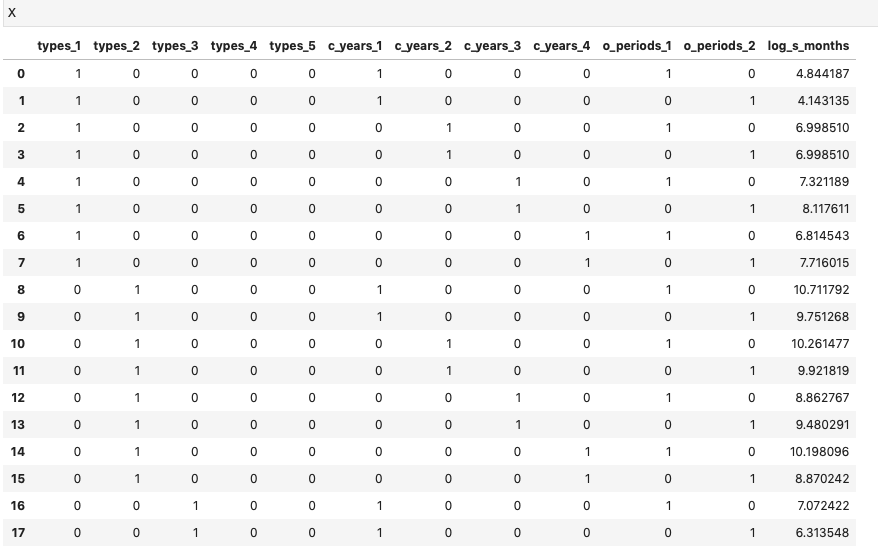
****

In addition, we input NumPy.log into both ‘ship’ datasets and ‘X’ datasets to print additional dataset called ‘log\_s\_months’;

****

****

Continue;

****

**Question 1c.**

There are too few rows to do a train test split in this dataset. In addition, as this is a regression problem on the number of incidents, the large standard deviation in the number of incidents may make the prediction even more inaccurate.

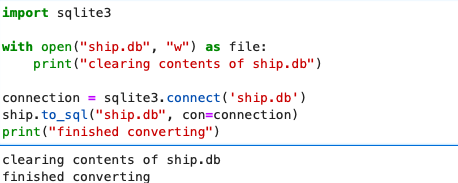
**Question 1d.**

Using the file from question 1b, we will save the file from ‘ship’ to ‘ship\_prepared’ using pandas.Dataframe.to\_csv (pandas, 2021)

****

****

To convert ships.csv to ship.db we will import SQL query (stack overflow, 2021) into the text in order to convert accordingly;

****

the programme will appear in the files as shown below:

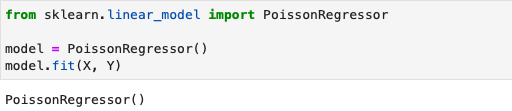
****

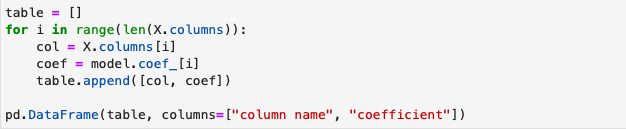
**Question 2a.**

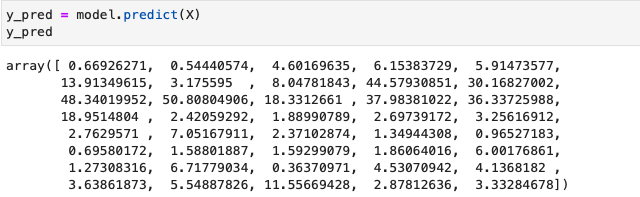
The module is Poisson Regressor, and is found in sklearn.linear\_model. Module: linear\_model, which contains other models for example LinearRegression, LogisticRegression. Estimator: PoissonRegressor. Fit function: This function trains the model using the training dataset. It takes in a x (predictor) and y (target) variables. Predict function: This function takes in and x (predictor) variable, and returns us a prediction

**Question 2b.**

In this question, we will import poisson regressor (scikit-learn, 2021) to solve the following equation from dataframe X and Y given in question 1

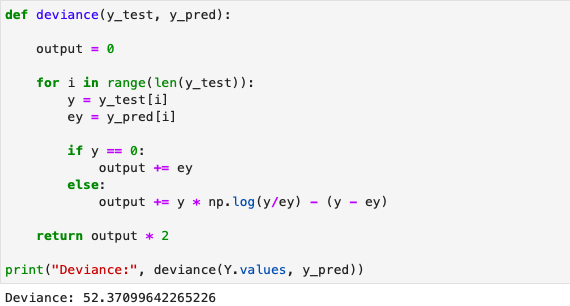
****

****

****

**Question 2c.**

****

****

**End of assignment**

# Bibliography

Yildirim, S. (2020, Feb 12). *Towards data science*. Retrieved from Handling Missing Values with Pandas A complete tutorial on how to detect and handle missing values : https://towardsdatascience.com/handling-missing-values-with-pandas-b876bf6f008f

pandas. (2021). *pandas*. Retrieved from pandas.DataFrame.info DataFrame.info(verbose=None, buf=None, max\_cols=None, memory\_usage=None, show\_counts=None, null\_counts=None)[source]: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.info.html

pandas. (2021). *pandas.get\_dummies*. Retrieved from pandas.get\_dummies(data, prefix=None, prefix\_sep='\_', dummy\_na=False, columns=None, sparse=False, drop\_first=False, dtype=None)[source] Convert categorical variable into dummy/indicator variables.: https://pandas.pydata.org/docs/reference/api/pandas.get\_dummies.html

pandas. (2021). Retrieved from https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.to\_csv.html

stack overflow. (2021, Aug). *stach overflow*. Retrieved from Importing a CSV file into a sqlite3 database table using Python Ask: https://stackoverflow.com/questions/2887878/importing-a-csv-file-into-a-sqlite3-database-table-using-python

scikit-learn. (2021). *sklearn.linear\_model.PoissonRegressor*. Retrieved from Generalized Linear Model with a Poisson distribution. This regressor uses the ‘log’ link function. Read more in the User Guide. New in version 0.23.: https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.PoissonRegressor.html

Wu, K. Y. (2021). *ANL252 Python for data analytics (study guide)*. Singapore University of Social Sciences. Release V1.0 Build S1.0.5, T1.5.21