

**Statistics for Data Analytics**

**CA-2**

**Prepared by: VARUN GOWDA**

**Student ID : 18129129**

**Submission date: April 7, 2019**

STATISTICS FOR DATA ANALYTICS – CV 2

ANALYSIS PART I Multiple Regression Model

ABSTRACT

New Zealand is a beautiful country, it has a lot to offer be it beaches or mountains, glaciers, lakes and volcanoes. The people are very welcoming, and the country has a great road network. The climate is perfect and pleasant all around the year. There is an exclusive dataset discussing The Accommodation Survey records guests staying in hotels, motels, backpacker accommodation, and holiday parks in New Zealand each month. The link for the dataset: <https://catalogue.data.govt.nz/dataset/accommodation-survey>. We are interested to know how occupancy rate in Holiday Parks is low in New Zealand in comparison to other facilities and try to fit a multiple regression model on it so that we can get an intuitive Idea as to what all variables impact Occupancy Rate.

DATASET

The dataset contains pivoted data from various months from 2001 to 2018 for 4 files i.e. Backpackers, Hotels, Motels, Holiday Park (Hereafter referred to as park). There are 11 Main fields in each file:

* Number of establishments
* Daily capacity (stay-units available)
* Monthly capacity (stay-unit nights available)
* Occupancy rate (%)
* Guest nights
* Guest arrivals
* Stay-unit nights (occupancy)
* Average length of stay (days)
* Guests per stay-unit night
* Stay-units per establishment
* Guest night % of January guest nights

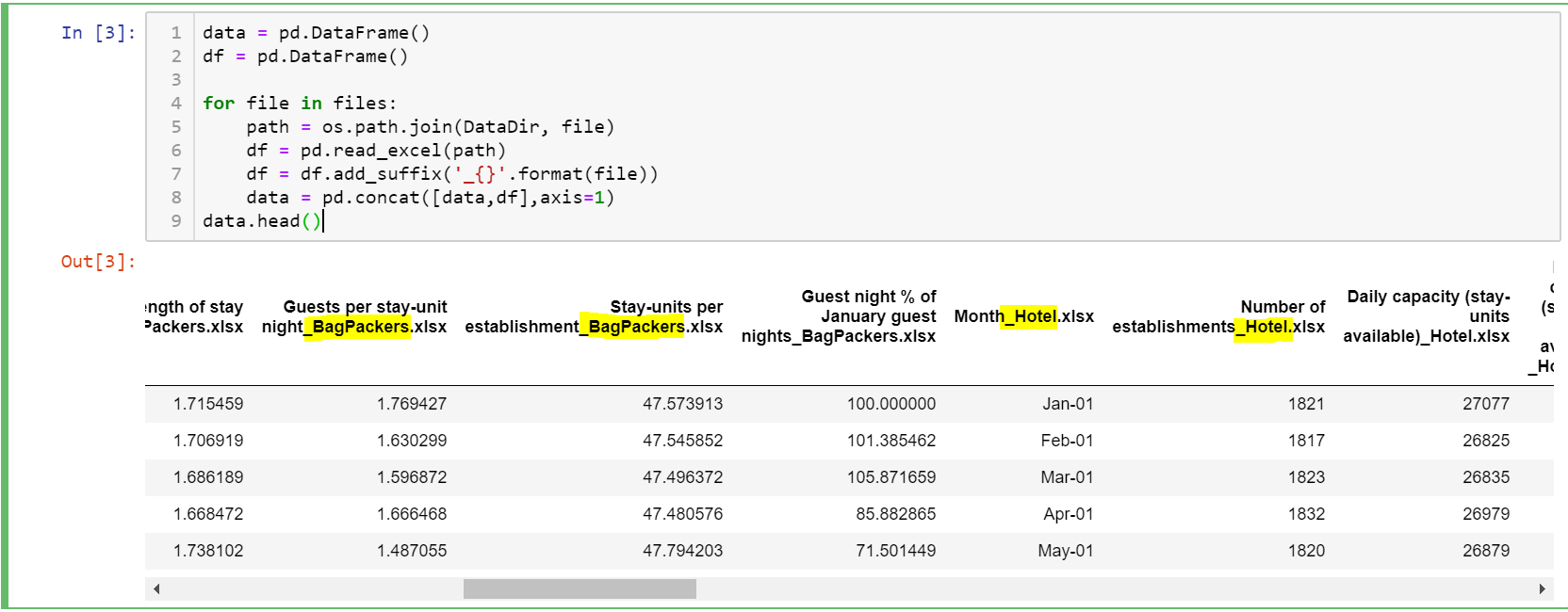
OBJECTIVE OF ANALYSIS

Our objective of analysis is to find which variables influence the Occupancy rate (%) of Holiday Parks

and what can we do to lessen the shortage in occupancy or if need be remove those and set up other more occupied hotels/motels. I have used python to run the analysis.

DATA CLEANING AND TRANSFORMATION

DATA IMPORT: The first step here is to join the four files with the same column. Here the challenge is that it is time series data and it cannot be clubbed one under the other. So, I first filtered the 4 Accommodation types from the data file and then created separate excel files for each. We used the following code to get entire data under one roof:



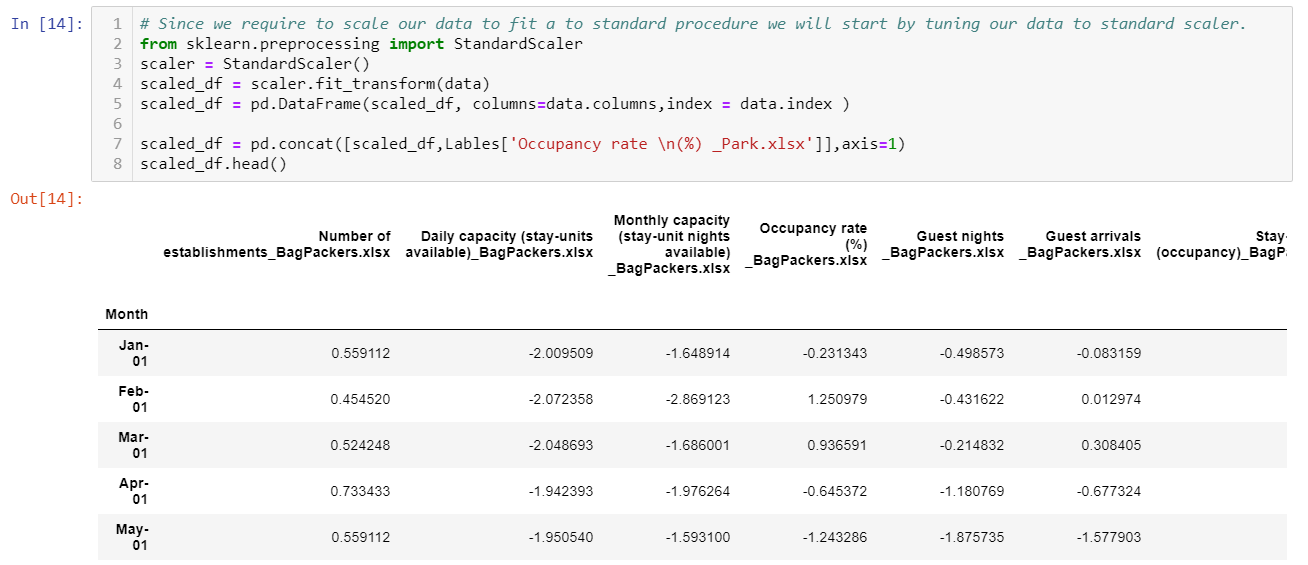
REMOVING Unnecessary columns : Here we have removed the redundant columns like Month Hotel, Month Motel and Month Park as we have already indexed by Month\_Bagpacker as an Index.

CHECKING for Missing Values : We have checked for missing values in the dataframe and we find that there is no missing value.

PREPROCESSING

We have scaled the columns using standard scaler to bring all the columns to the same scale as there are some percentage columns too. We have also used the Heatmap to find the correlation between the variables in the dataframe.

Standard Scaler

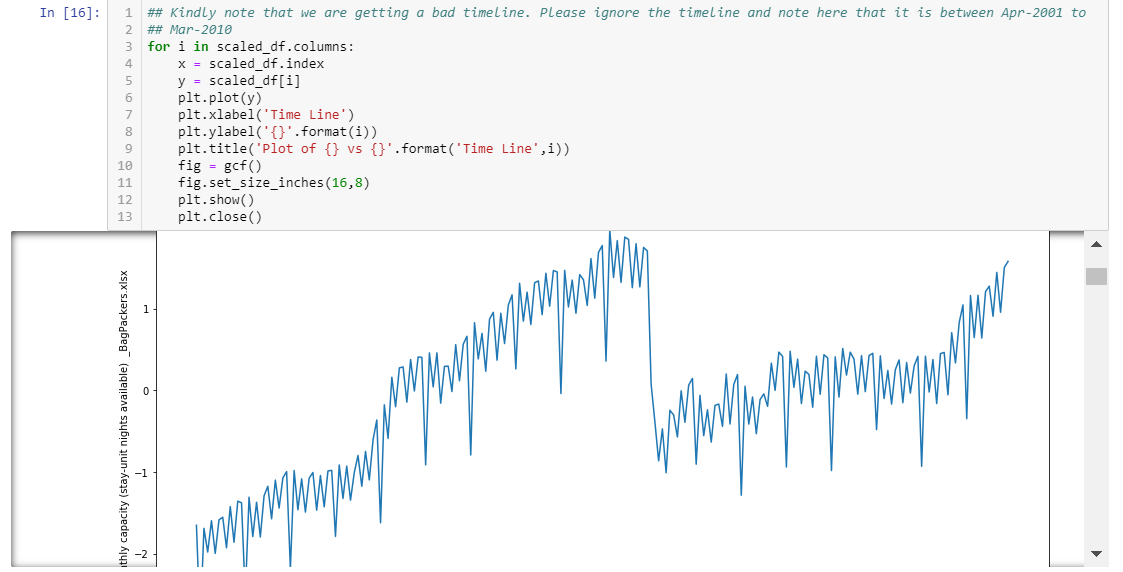


HEAT MAP:

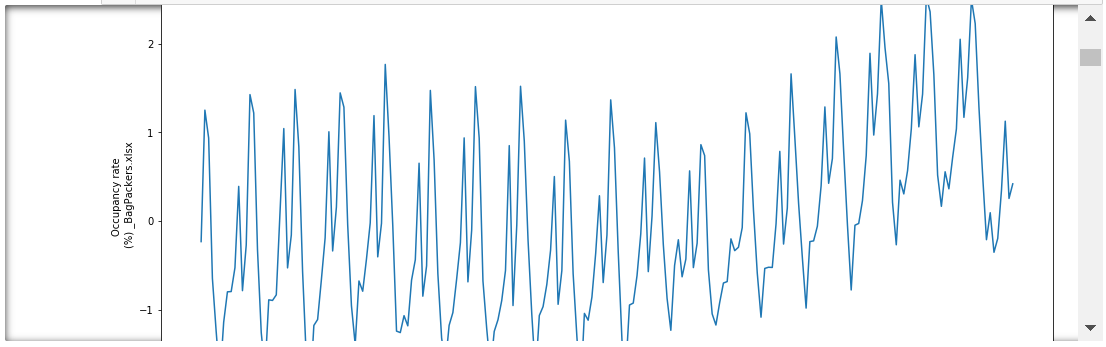
A **correlation heatmap** uses colored cells, typically in a monochromatic scale, to show a 2D **correlation** matrix (table) between two discrete dimensions or event types. ... **Correlation heatmaps** are ideal for comparing the measurement for each pair of dimension values.



As a part of time series analysis, we also need to check for the variables if they are stationary or not. Based on this we check for a Zero Mean Condition. I have plotted graphs for this case:

Example: This Variable doesn’t contain seasonality

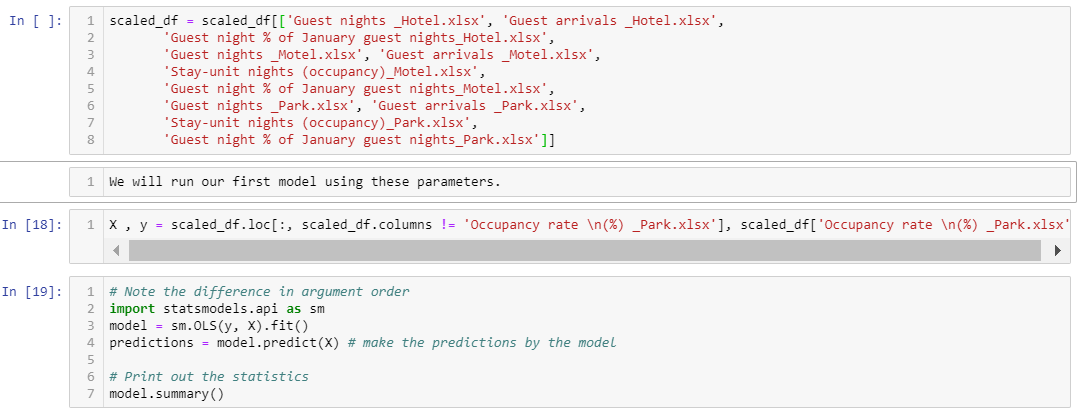
This variable contains seasonality:



Upon Checking for all these conditions, we can say with enough confidence that our assumptions are correct given the parameters.

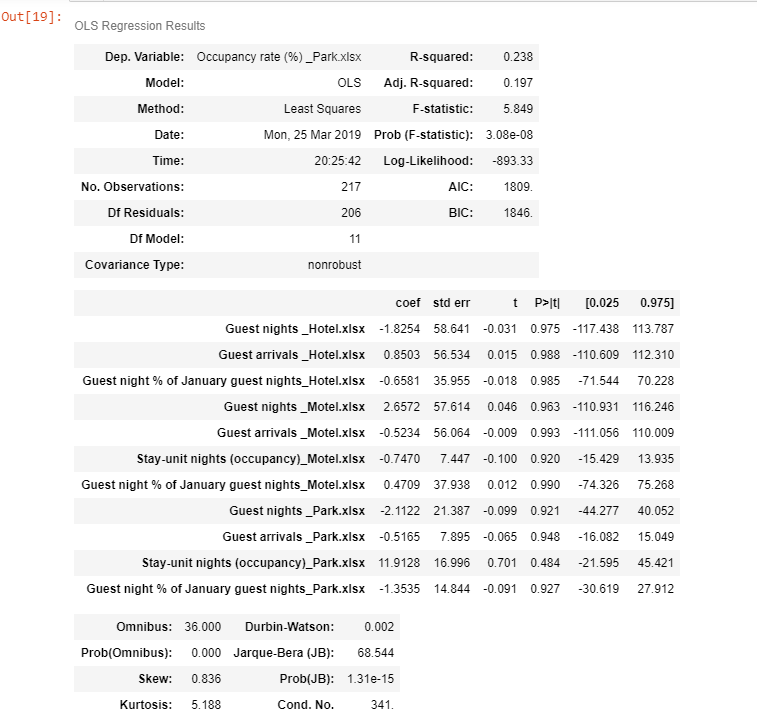
MODELING

We now split the data into Features and Labels and then run a Multiple Linear Regression Model for this case. Following code snippet discusses the partitioning into features and Labels followed by running the model on the feature extracted columns.

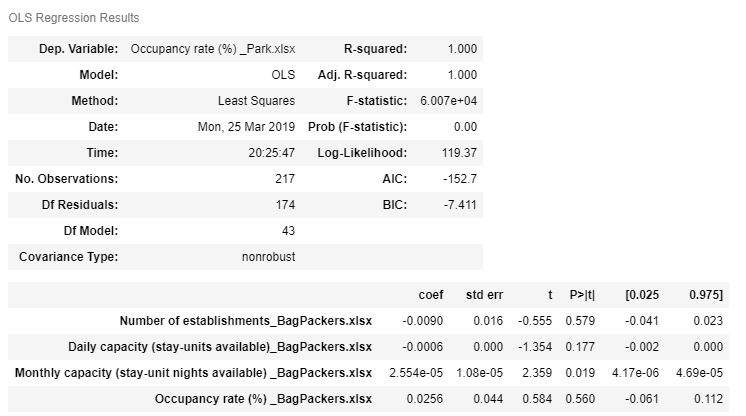


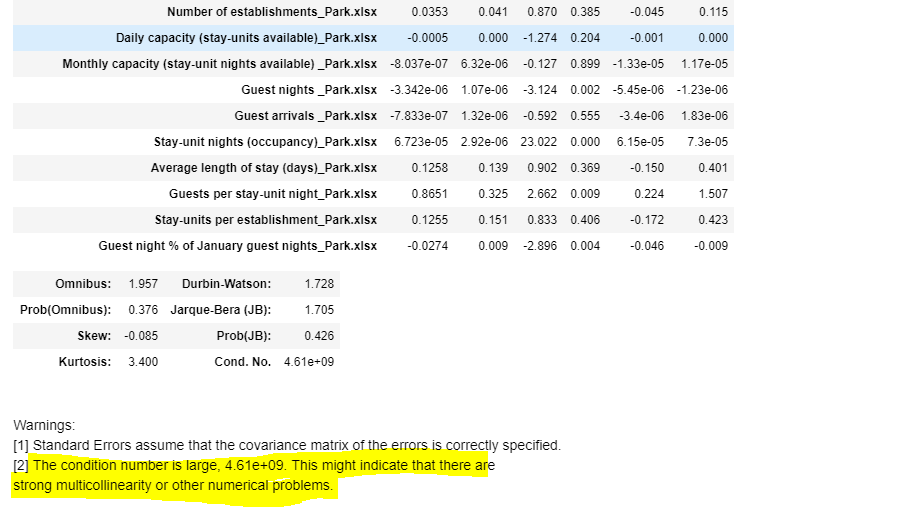
I also ran the model on all the feature columns without removing any column. The reason the second model is flawed because there is auto-correlation although the R squared value is 1.

MODELING: MODEL ON SCALED DATAFRAME



This model suggests that Guest Nights and Guest Arrivals are the parameters which are closely related to Modeling the Occupancy Rate. But we can only accept this at 5 % significance level.





We are sure that this model is not that perfect in respect of the columns that we have chosen as we are aware that this is time dependent data. Our intuition about the columns is inappropriate in the sense that this data is time series data. From the above model we get that there is strong multi collinearity because there is an autocorrelation between the different timeseries values.

ANALYSIS PART II Binary Logistic Regression Implementation

ABSTRACT

There is an Interesting Dataset on Bio Security of Current and Past Plants. The dataset discusses various Plant names, scientific names and lists the date when it was found and many more such attributes. It appears that the Status is a binary level of Current and Past. This motivated me to study their relation within the ambit of Logistic Regression. The link for the dataset is: https://catalogue.data.govt.nz/dataset/biosecurity-plant-pest-containment.

DATASET

The dataset contains the following attributes :

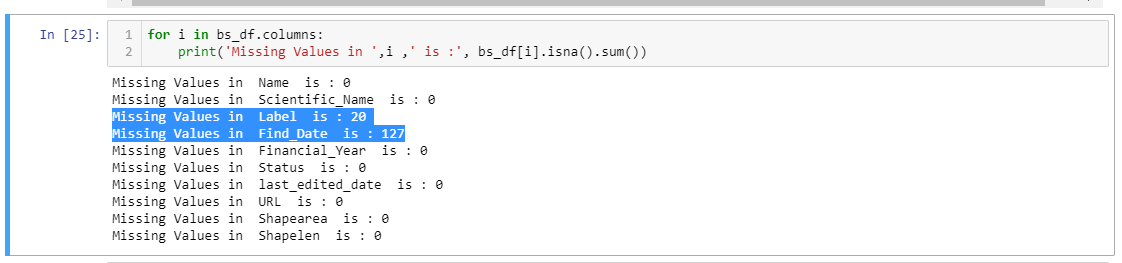
* Name
* Scientific\_Name
* Label
* Find\_Date
* Financial\_Year
* Status
* last\_edited\_date
* URL
* Shapearea
* Shapelen

OBJECTIVE OF ANALYSIS

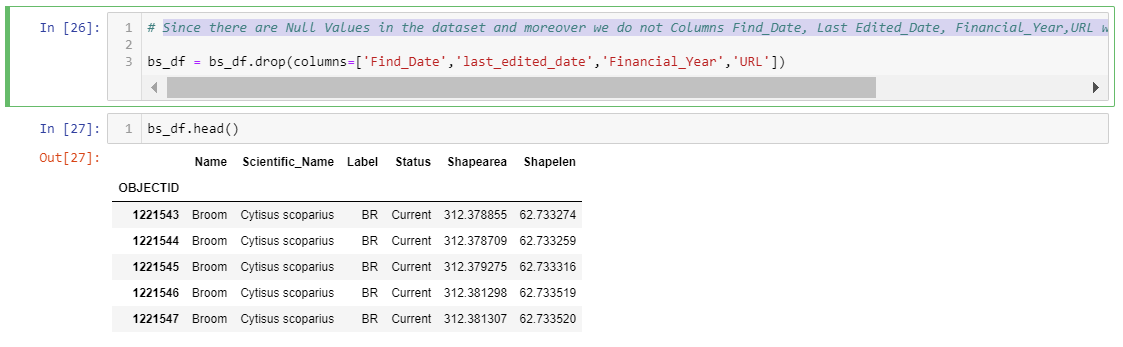
Our objective of analysis is to find which variables influence the Status of the Plant to be classified into Current and Past status. We will eventually print a confusion matrix to predict and state what the Accuracy of our model is. Python is used to run the analysis.

DATA CLEANING AND TRANSFORMATION

CHECKING FOR MISSING VALUES : We have checked for missing values in the dataframe and we find that there are missing values in Label and Find\_Date Columns.



REMOVING COLUMNS : Since we see that there are null values in Label and Find\_Date we need to check for it. We removed the Find\_Date column, Although Find\_Date is a good parameter to check for current and Past dates we have not taken it, since this tends to give us a trivial solution by partitioning based on date column. We go on to remove 'Find\_Date','last\_edited\_date','Financial\_Year','URL' further.



REMOVING LABEL N/A VALUE COLUMNS : Next Label is another variable which is crucial in assigning the type of plant it is. So, we keep this variable and remove the 20 rows having missing Labels.

ONE HOT ENCODING: Next, we do one hot encoding for the categorical columns: One hot encoding is a process by which categorical variables are converted into a form that could be provided to algorithms to do a better job in prediction.

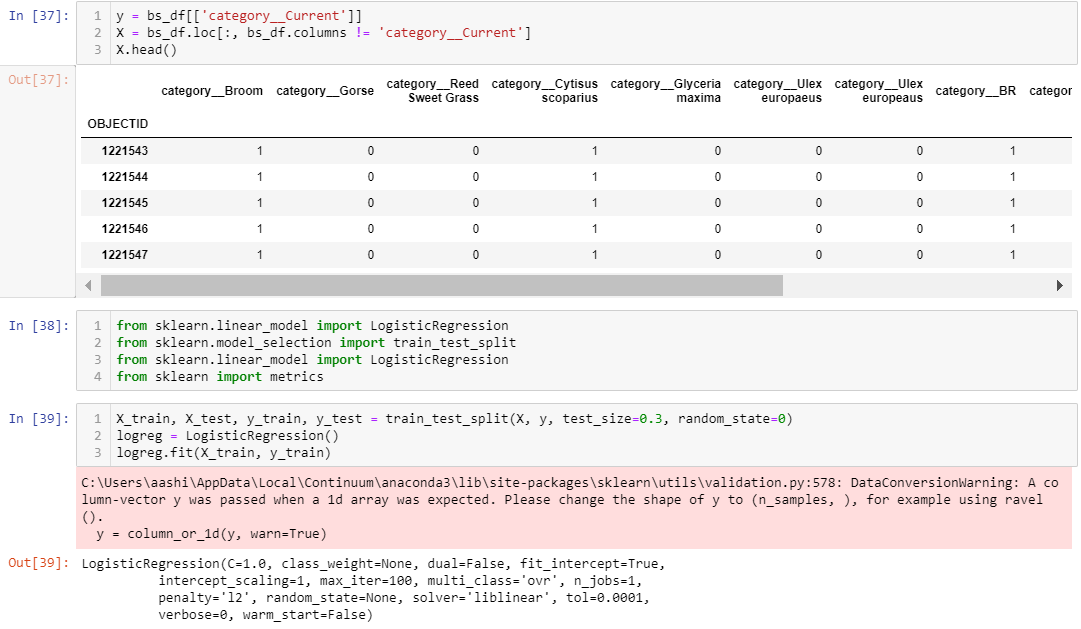


PREPROCESSING

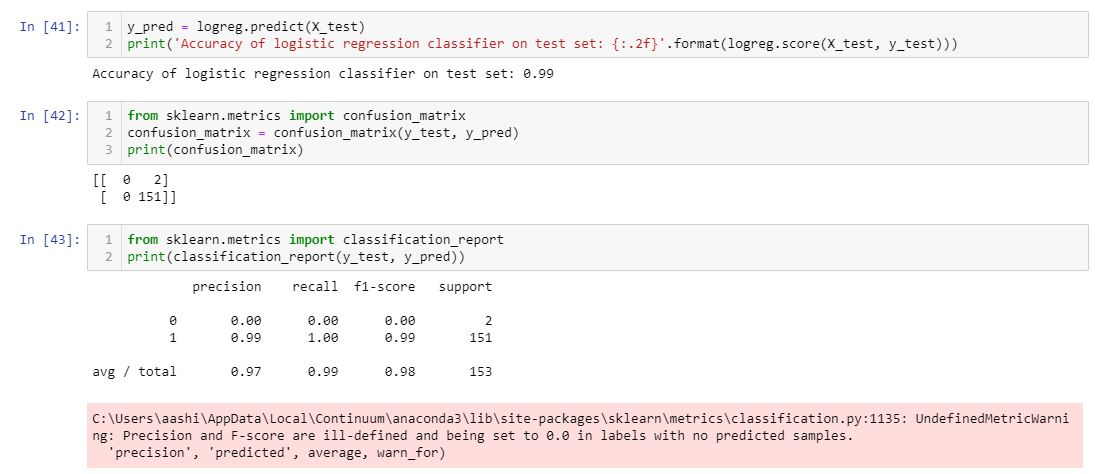
Now once we have concatenated the data we carry on with printing the correlation matrix to check for which variables are needed for modeling the data after we joined the two data sets.

As the data to be analyzed is for Status, we remove one of the columns so that we can remove the perfect correlation between category\_past and category\_current.

MODELING: We remove the category\_current column to get the feature dataset and append it to an empty Dataframe called y. I then imported the Logistic Regression Model from Sklearn and ran it through a sample test.



ACCURACY AND CONFUSION MATRIX



We here get the Regression accuracy of 99% with a false negative rate of 0.01 This proves that even without the dates our model based on Shape length and Shape area is able to categorize in Present and Past values.

**Reference:**

Catalogue.data.govt.nz. (2019). *Accommodation Survey - data.govt.nz - discover and use data*. [online] Available at: https://catalogue.data.govt.nz/dataset/accommodation-survey [Accessed 7 Apr. 2019].

Catalogue.data.govt.nz. (2019). *Biosecurity - Plant Pest Containment - data.govt.nz - discover and use data*. [online] Available at: https://catalogue.data.govt.nz/dataset/biosecurity-plant-pest-containment [Accessed 7 Apr. 2019].

Docs.arcadiadata.com. (2019). *Correlation Heatmap Visuals*. [online] Available at: http://docs.arcadiadata.com/4.0.0/pages/topics/visual-correlation.html [Accessed 7 Apr. 2019].