

# CA4: Pandas Data Interpretation – 10%

Module Title: Programming for Big Data

Module Code: B8IT105

Module Leader: Darren Redmond

Student Name: Karina Jonina

Student Code: 10543032

Github: <https://github.com/KarinaPS11/B8IT105>

Contents

[CA4: Pandas Data Interpretation – 10% 1](#_Toc41742779)

[1. Dataset 3](#_Toc41742780)

[2. Investigative Questions 4](#_Toc41742781)

[3. Number of Reservations 5](#_Toc41742782)

[4. Number of Reservations Per Month and Year 6](#_Toc41742783)

[5. Total Number of Guests 7](#_Toc41742784)

[6. Total Number of Nights Per Reservation 8](#_Toc41742785)

[7. Nationality of Visitors 9](#_Toc41742786)

[8. Top 10 Countries 10](#_Toc41742787)

[9. Reservation Status – Check-Outs, Cancelations and No-Shows 11](#_Toc41742788)

[10. Previous Cancelation 16](#_Toc41742789)

[11. Average Daily Rate Per Night in Each Month 18](#_Toc41742790)

[12. Average Daily Rate Per Night in Each Country 19](#_Toc41742791)

[13. Meal Types 20](#_Toc41742792)

[14. Matching Expectations and Reality 22](#_Toc41742793)

[15. Conclusion 23](#_Toc41742794)

# Dataset

The dataset used was collected from the following website:

<https://www.sciencedirect.com/science/article/pii/S2352340918315191>

This dataset was collected by two hotels. Both hotels are located in Portugal. The dataset has 31 variables with 40,060 observations for resort hotel in Algarve and 79,330 observations of city hotel in Lisbon. The dataset was collected between 1st of July 2015 and 31st August 2017.

The following are column names:

'hotel', 'is\_canceled', 'lead\_time', 'arrival\_date\_year', 'arrival\_date\_month', 'arrival\_date\_week\_number', 'arrival\_date\_day\_of\_month', 'stays\_in\_weekend\_nights', 'stays\_in\_week\_nights', 'adults', 'children', 'babies', 'meal', 'country', 'market\_segment', 'distribution\_channel', 'is\_repeated\_guest', 'previous\_cancellations', 'previous\_bookings\_not\_canceled', 'reserved\_room\_type', 'assigned\_room\_type', 'booking\_changes', 'deposit\_type', 'agent', 'company', 'days\_in\_waiting\_list', 'customer\_type', 'adr', 'required\_car\_parking\_spaces'

# Investigative Questions

The following are the questions that will be answered

* Most popular months and years
* Number of guests per reservation
* Number of nights per reservation
* Nationality of guests
* Price of a room per person per night
* Number of cancellations based on nationalities and meal types
* Nationalities with the greatest number of reservations
* Month and Year with the highest number of cancelations

# Number of Reservations

A screenshot of a cell phone

Description automatically generated

Graph 1: It is clear from the graph that August, July and May are the Top 3 months for travel. The winter months are the three least popular months for travel.

# Number of Reservations Per Month and Year

A screenshot of a cell phone

Description automatically generated

Graph 2: This graph shows trend reservations by month and year. There is a high number of reservations every year during spring and summer. As expected, there are reservations decrease in winder months every year. Furthermore, there has been a greater number of reservations every year.

# Total Number of Guests

Graph 3 Data Preparation

There were three individual columns for 'adults', 'children', 'babies'. Total number of guests was calculated by adding the three columns together.

A screenshot of a cell phone

Description automatically generated

Graph 3: The trend shows that majority of reservations are for two people. The second highest trend is for just one person. It can be presumed that these are solo travellers.

# Total Number of Nights Per Reservation

Graph 4 Data Preparation

There were two individual columns for 'stays\_in\_weekend\_nights' and 'stays\_in\_week\_nights'. Total number of nights was calculated by adding the two columns together.

A screenshot of a cell phone

Description automatically generated

Graph 4: The trend shows that majority of reservations stay for two or three night. 'It can be presumed that this is reflected by people going away by Friday, Saturday and Sunday.

# Nationality of Visitors

Graph 5: Graphs shows that the most popular nationality of the visitors is Portuguese. However, there is a big droA screenshot of a cell phone

Description automatically generatedp to the second most popular nationality. The bars slowly decrease after the second most popular nationality. Reservations from Portugal came to 48590, while reservation from Norway were 607.

# Top 10 Countries

Graph 6 Data Preparation

Due to the large amount of countries (178) and the numbers of reservations for those countries were tiny, it was decided to investigate data in Top 10 countries. Each country in the data was split from the data and the merged using the ‘concat’ function in Python.

A picture containing umbrella

Description automatically generated

Graph 6: The Top 10 countries were filtered from the data, the breakdown that Portuguese makes up nearly half of the data. British and French tourists share less than an eighth each. The rest of the countries are less than 10% each.

# Reservation Status – Check-Outs, Cancelations and No-Shows

A close up of a logo

Description automatically generated

*Graph 7:* Less 2/3 of reservations get fulfilled. Nearly 40% of reservations are cancelled or there is a no-show. Don’t these people want to go on a holiday?

A picture containing implement, pencil

Description automatically generated

Graph 8: Using Graph 2, check-outs and cancelations were analysed. Spring and summer months are most popular for travel. This reflects in cancelations as well.

A screenshot of a cell phone

Description automatically generated

Graph 9: As shown in Graph 5, Portuguese nationals make the greatest number of reservations. Unfortunately, it the number of cancelations is greater than the number of people checking-out for the Portuguese.

Graph 10 Data Preparation

There is a column in the data called 'is\_cancelled’. The number [1] is assigned to cancelations while [0] is assigned to check-outs. Sub-set of the data was created with only cancelations.

A picture containing umbrella

Description automatically generated

*Graph 10*: Portuguese reservations make up 70% of all cancelations, while the other 9 nationalities make up the other 30% of cancelations in Top 10 Countries.

Graph 11 Data Preparation

When each country in the Top 10 was split from the data and the merged using the ‘concat’ function, a separate dataset for Portugal was made. This dataset was used to analyse cancelations in Portugal.

A pencil and paper

Description automatically generated

*Graph 11*: Follow up from *Graph 8*, Portuguese nationals’ cancelation for each month and year were examined. Portuguese nationals check-out rate had a drop every spring and summer months in 2016, while their cancelation increased for the same months.

# Previous Cancelation

Graph 12 Data Preparation

There is a column in the data called 'is\_cancelled’. The number [1] is assigned to cancelations while [0] is assigned to check-outs. Sub-set of the data was created with only cancelations. The column 'previous\_cancellations’ was also used to create this graph. Due to the high number of reservation that did not have a previous cancelation, values of greater than 1 were used. Otherwise the graph would be skewed.

Table returned:

* 33 245 reservations had **never** made a cancelation before.
* 5 685 reservations had made **1** previous cancelation.
* 48 reservations had made **24** previouscancelations
* 37 reservations had made **2** previouscancelations.
* 26 reservations had made **26** previouscancelations.
* 25 reservations had made **25** previous cancelations.
* 20 reservations had made **3** previous cancelations.
* 19 reservations had made **19** previous cancelations.
* 14 reservations had made **14** previous cancelations.
* 11 reservations had made **13** previous cancelations.
* 10 reservations had made **11** previous cancelations.
* 7 reservations had made **6** previous cancelations.
* 7 reservations had made **4** previous cancelations.
* 2 reservations had made **5** previous cancelations.
* 1 reservations had made **21** previous cancelations.

A screenshot of a cell phone

Description automatically generated

*Graph 12*: Who knew that there were that people who are repeatedly cancel holidays?

# Average Daily Rate Per Night in Each Month

Data Preparation

The data had ‘adr’ column. To calculate the price of the room per person per night, the following formula was devised: ‘adr’ / ‘total\_guests’

A screenshot of a cell phone

Description automatically generated

Graph 13: The prices reflect the most popular trend. The late spring and summer months are some of the most expensive months.

# Average Daily Rate Per Night in Each Country

A screenshot of a cell phone

Description automatically generated

Graph 14: Graph shows that there is just a small difference in price between countries. Although, in a previous graph, we saw that there were more cancelations and no-shows for Portuguese nationals.

# Meal Types

There are 77 824 Bed and Breakfast (BB), 13 297 Half-Board (HB), 7781 Self-Catering (SC), 1 115 Undefined and 783 Full-Board (FB).

A close up of a logo

Description automatically generated

Graph 15: Graph shows that over 75% of the data was the Bed ‘h Breakfast hotels. Half-Boards make up just over an eight of the data.

A screenshot of a cell phone

Description automatically generated

Graph 16: Bed ‘n Breakfast have the most reservations, but also have the most number of cancelations.

# Matching Expectations and Reality

Graph 16 Data Preparation

There are two columns in the data called 'reserved\_room\_type' and 'assigned\_room\_type'. A boolean column was created with ‘False’ and ‘True’, meaning that the reserved room matched the assigned room. ‘False’ meant that the assigned room (Reality) did not ‘match’ the reserved room (Expectations).

A close up of a logo

Description automatically generated

*Graph 17:* Over one in eight reserved rooms did not match their assigned room.

# Conclusion

The given dataset examines the cancelation rate of two hotels in Portugal, one is a hotel in Lisbon and one hotel is a Resort in Algarve. The graphs painted an interesting model.

The findings suggest that:

* Most people come in together in groups of two and three.
* However, the hotels are popular with solo tourists.
* Most people stay two to three nights.
* Most visitors are from Portugal, Great Britain, Spain, Germany and Italy.
* Hotels have been growing in number of reservations every year.
* The most popular seasons are always spring and summer.
* Following the two above observations, this also means that there are more cancelations every year and there are more cancelations during the summer.
* Prices rise for the spring and summer months.
* 70% of cancelation are from Portugal.
* Most popular meal type is Bed and Breadkfast, however this is also reflected in its cancelation rate.
* One eight of the bookings are not met with their reserved room type.

Machine Learning will be addressed in CA5.