LUMEN DATA SCIENCE 2024.

PROJECT DOCUMENTATION

PREDICTION OF HOTEL OCCUPANCY

Matija Kukić, Ivan Džanija, Valentina Vidović, Dora Baričević

Zagreb, April 2024

Table of Contents

1. General overview

1.1. Introduction

In the competitive landscape of the hospitality industry, understanding and predicting hotel occupancy rates is important for efficiency, revenue management and making sure guests are satisfied. As a part of a student competition we have been tasked with solving this problem. We have embarked on a project to develop a predictive model capable of forecasting hotel occupancy rates with high accuracy.

1.2. Objective

This documentation outlines the methodologies, data analyses and technological tools utilized in our predictive modeling project. Our primary goal is to enable the hotel management to make informed decisions based on accurate occupancy forecasts, thereby optimizing resource allocation, pricing strategies and improving the overall guest experience.

1.3. Project overview

The project leverages a dataset for a Croatian hotel comprising historical reservation data ranging from 2008 to 2009, containing data on reservation id, stay dates, country and so on. By applying data analytics and machine learning techniques, we want to uncover patterns and dependencies that affect occupancy rates. This document details each step of our analytical process, from data preprocessing and exploration to model development and validation.

2. Project plan

1. Brief overview of the given dataset, preprocessing and initial transformations.
2. Processing the data and adding some obvious features to help the further data exploration.
3. Clean up the data set. Removing all the duplicate reservation, arrival date, departure date, cancelation date, adult count errors.
4. General analysis. Finding out types of resorts, unique types of reservation status, frequencies of guests based on their country, room type, adult count, children count, reservation length and more.
5. Deeper analysis. Analyzing the distributions of prices, room occupancy and room category.
6. Removing all other outliers that are impacting the prediction based on wrong data.
7. Finding and checking for correlations between the data.
8. Finding the parameters and training the model.
9. Making the point predictions and adding conformal prediction intervals.
10. Adding the prediction to the final application.

3. Requirements

For this project we will be using Python as our programming language. Also all the basic data science libraries for Python such as pandas and numpy, math and statistics. For our prediction models we chose ARIMA, but have also tried prophet and tensorflow's lstm.

If you wish to install all the dependencies we have described how to do it in README file.

4. Processing and cleaning

4.1. Loading data

The first step involved loading the hotel reservation data from provided parquet file, ensuring a comprehensive dataset is ready for cleaning and manipulation plus adding features such as nights spent and total cost.

4.2. Removing inconsistencies in dates

Pre-Reservation data: Entries where the arrival date was before the reservation creation date were removed, as this indicated a chronological error.

Post-Cancellation data: Entries with a reservation creation date after the cancellation date were also removed, suggesting data entry errors.

4.3. Filtering relevant data

Adult Guests: Data entries where the number of adult guests was zero were deleted, as these do not contribute to occupancy needs and predictions.

Cancellation Date Issues: Any records where the cancellation date was after the checkout date were removed to ensure accuracy in the data regarding actual stays.

4.4. Deduplication

The dataset included multiple entries for each guest's stay, split by day. These were consolidated into unique records per stay, utilizing the check-in and check-out dates to calculate the duration. This step was crucial as it reduced data redundancy and streamlined the dataset for further analysis.

5. Data analysis

We made a comprehensive analysis of the most important variables: number of guests, county, reservation status, room type and other.

5.1. General data overview

5.1.1. Reservation status

A visualization confirms the expected distribution of reservation statuses, indicating no apparent issues. Most of the reservations were successful and the quests Checked-out.

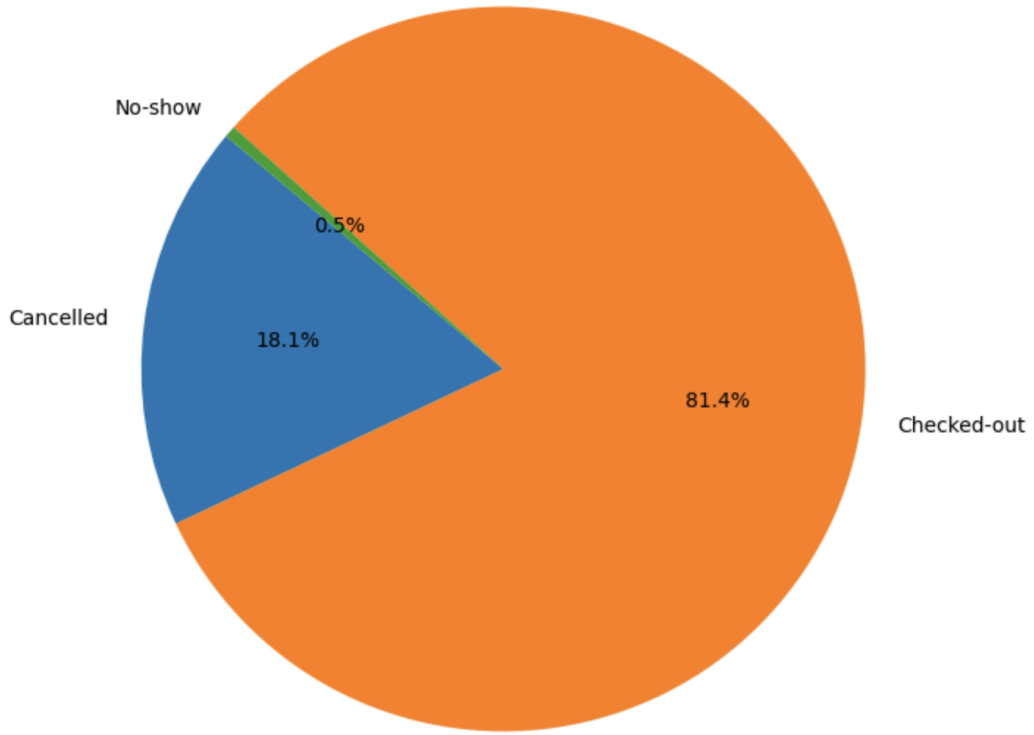

Figure 5.1.

5.1.2. Guest country

Analysis of guest origin shows no significant data concerns. Most of the guests came from Croatia, Italy, France, Great Britain, Netherlands and Slovenia. We will use this data in the future when setting up events time table that will contain the non-working days and the holidays of the relevant countries.

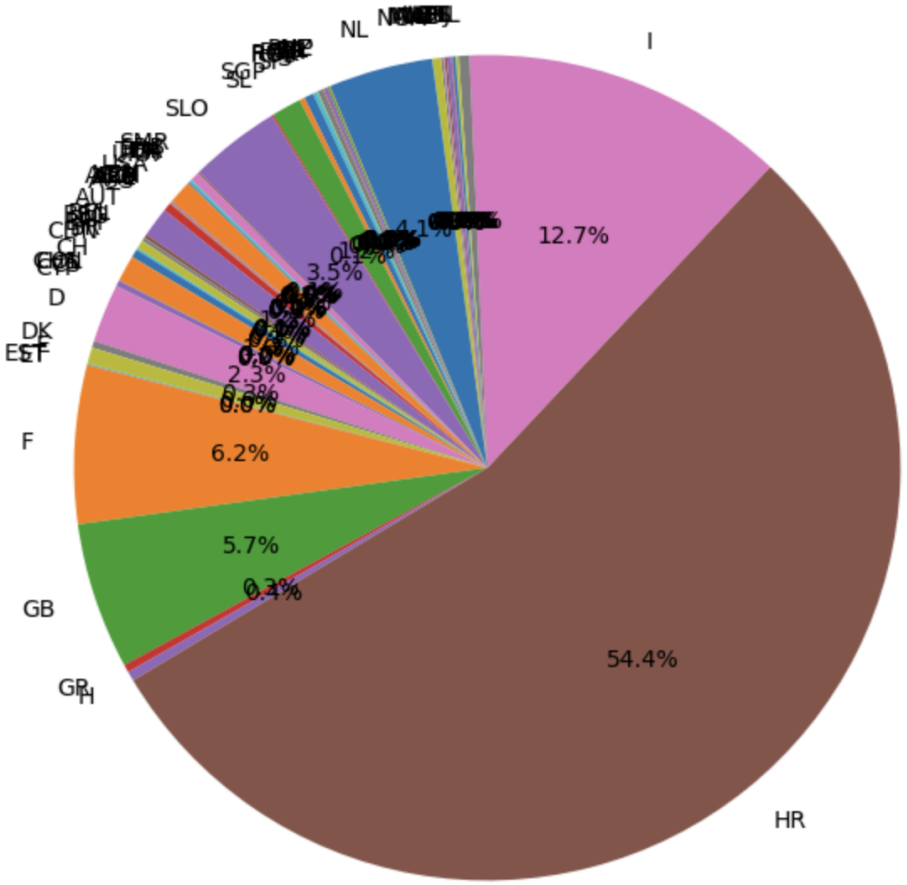

Figure 5.2.

5.1.3. Hotel ID

The data is confirmed to come from a single hotel or resort, ensuring consistency in the dataset.

5.1.4. Room category

Most reservations were made on rooms 5 and 2 indicating that this could be rooms for single person and/or cheaper rooms.

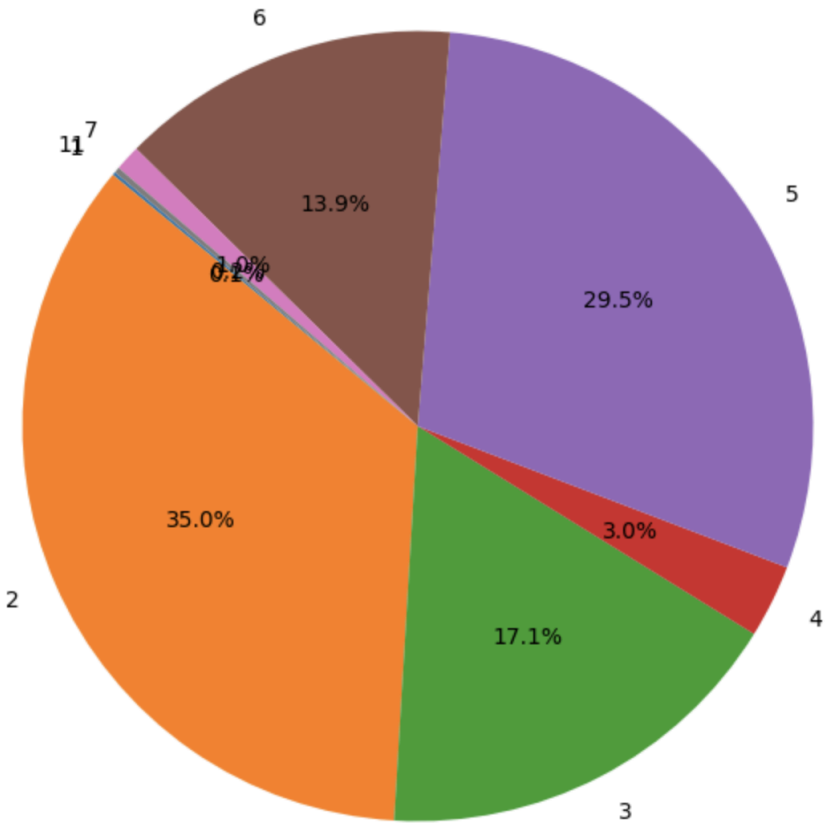

Figure 5.3.

5.1.5. Number of children

Identified as potentially anomalous due to unexpected values, leading to its exclusion in further analyses. The data showed only one reservation in two years had any children.

5.1.6. Number of adult guests

The distribution is as expected and shows no indication of data issues. Most of the reservations were made for one or two people which leads us to believe rooms 5 and 2 are single or double bedrooms.

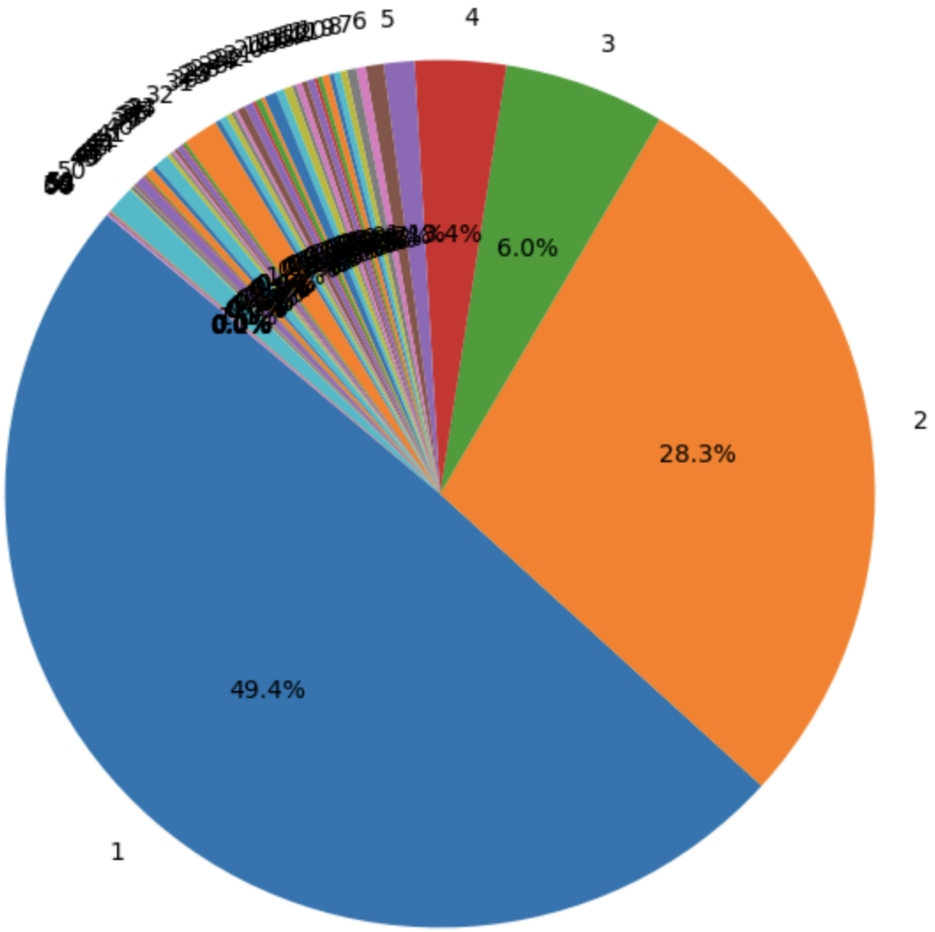

Figure 5.4.

5.1.7. Number of nights stayed

The analysis of the stay length shows reasonable and expected patterns. Most of the reservations were made for one night (62,5%) or for 2 nights (20,1%). We found outliers in the number of nights with the maximum of over 200 days.

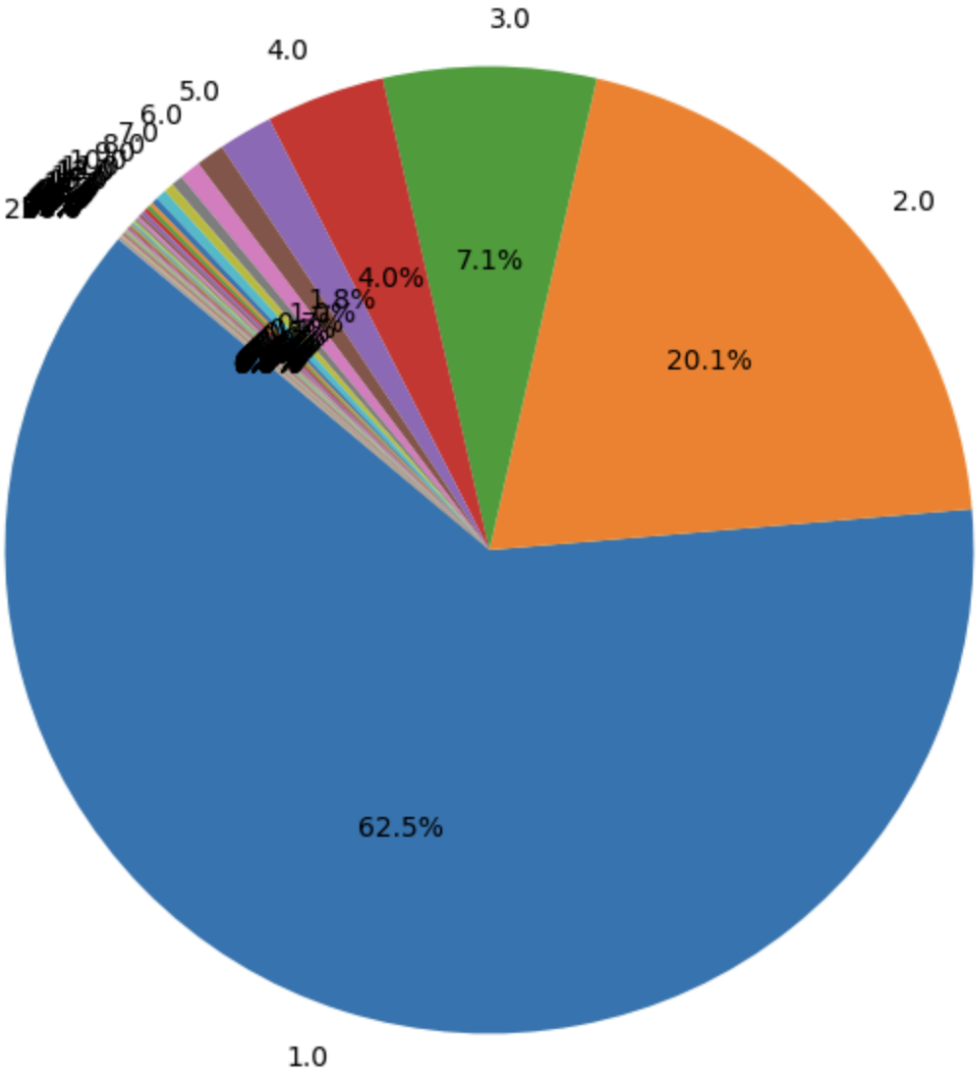

Figure 5.5.

5.2. Detailed data analysis

The focus shifts to reservations with a status of "Checked-out," analyzing the daily number of guests.

5.2.1. Daily guest count

The data is separated by days, and a total guest count per day is calculated and visualized to observe the distribution over time.

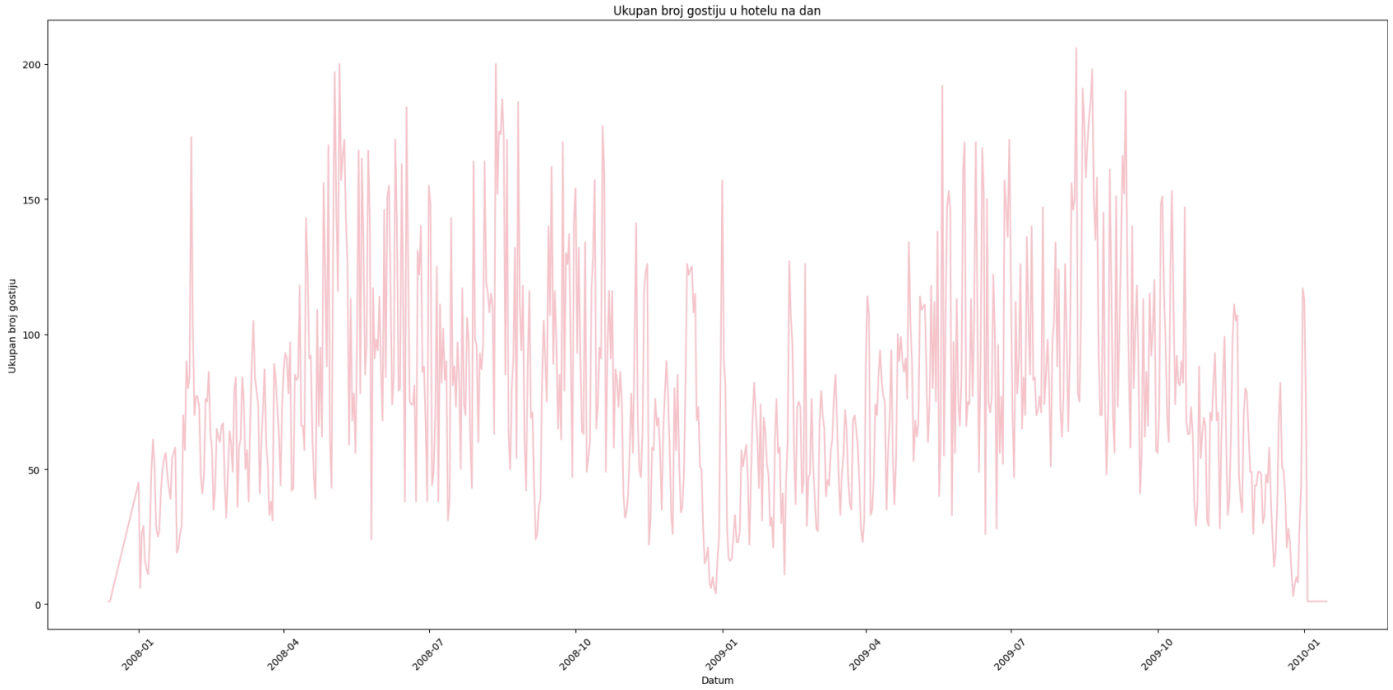

Figure 5.6.

The occupancy graph reveals notable dips in the first month, which are flagged for further investigation. It also shows expected similarities with corresponding months in previous years, suggesting potential seasonal patterns that could be explored further.

The data's volatility suggests that simpler predictive models might not provide pinpoint accuracy, emphasizing the need for establishing a reliable prediction interval instead of exact forecasts.

5.2.2. Daily reservation count

The data is again separated by days, and a total reservation count per day is calculated and visualized to observe the distribution over time.

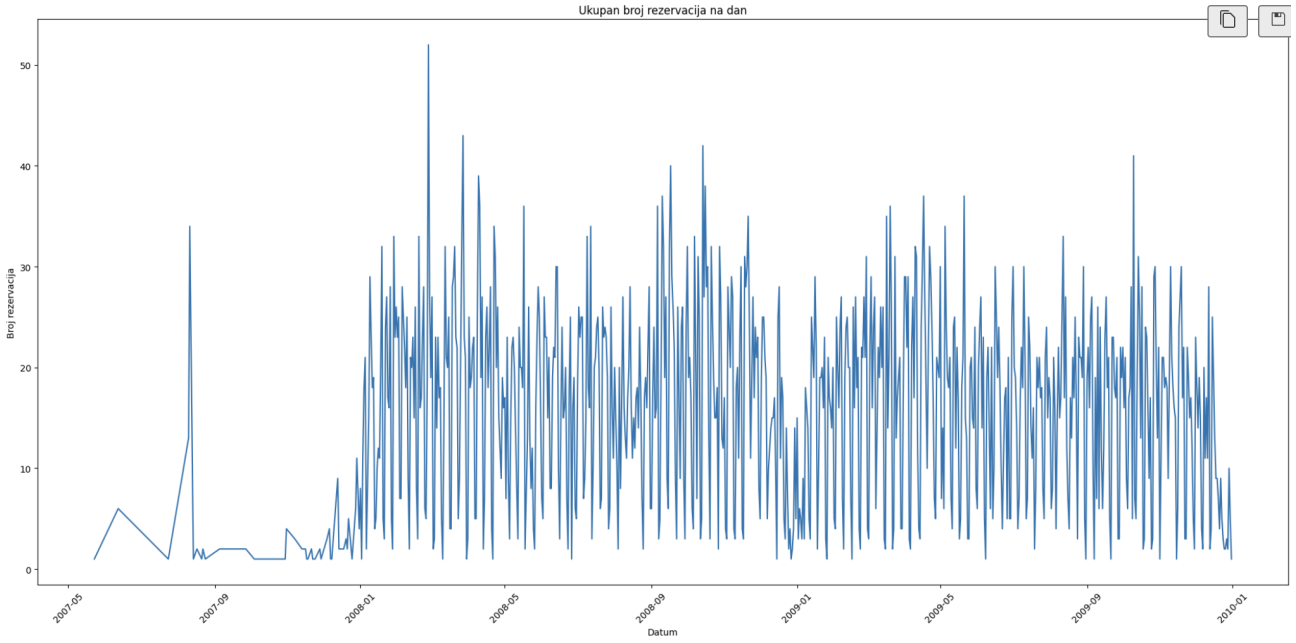

Figure 5.7.

The initial dip and lack of data is expected as some reservations are made months in advance but our data set is oriented towards arrival and departure dates. This graph also shows smaller variance so this could be a better prediction variable then daily guest count.

5.2.3. Daily occupied rooms count

The data is again separated by days, and a total room occupancy count per day is calculated and visualized to observe the distribution over time.

This is the variable of prediction we choose because it has smaller variance then the daily guest count but also better represents the goal of prediction then daily reservation count.

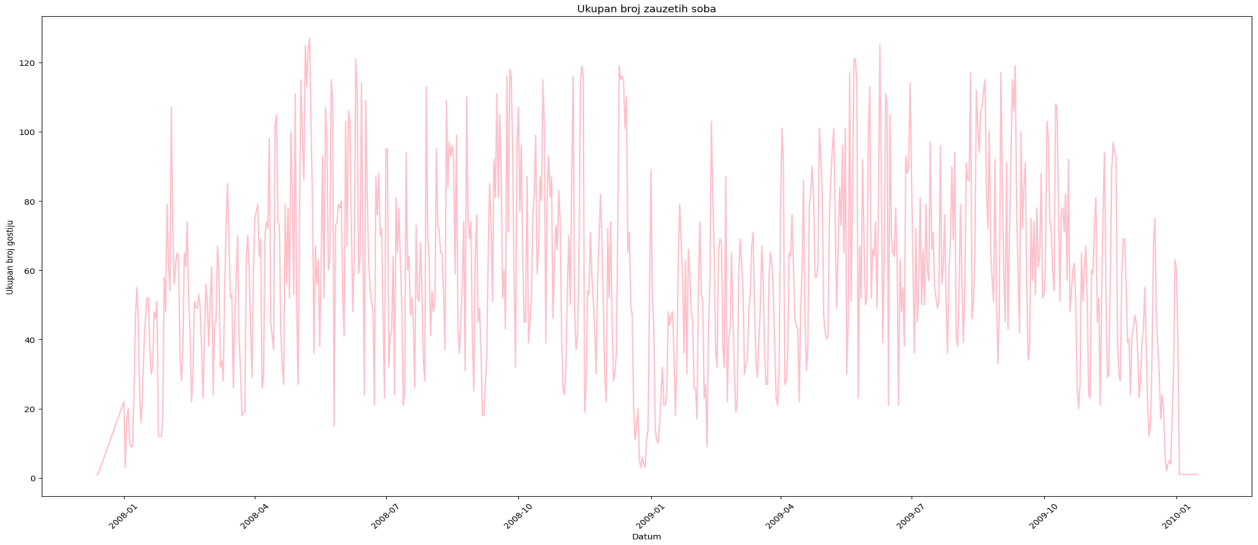

Figure 5.8.

5.3. Outlier removal

Here we tried to keep as much data as possible to avoid our model being overfit on the training data and also be able to guess minor variances. We only removed the reservations that had spent more than 6500 HRK per night of stay because that price was outrageous for that time period and also for the location.

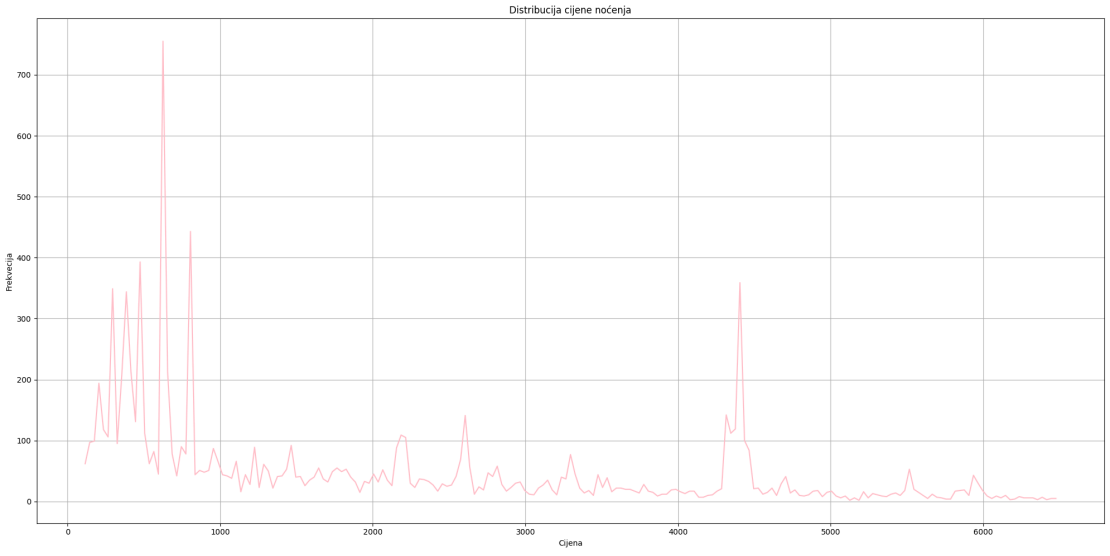

Figure 5.9.

This is the final graph of prices per night and this is an acceptable result.

5.4. Correlations between data types

We have found very little correlation with the date of the stay and any other data type and this was favorable for our simple prediction model ARIMA.

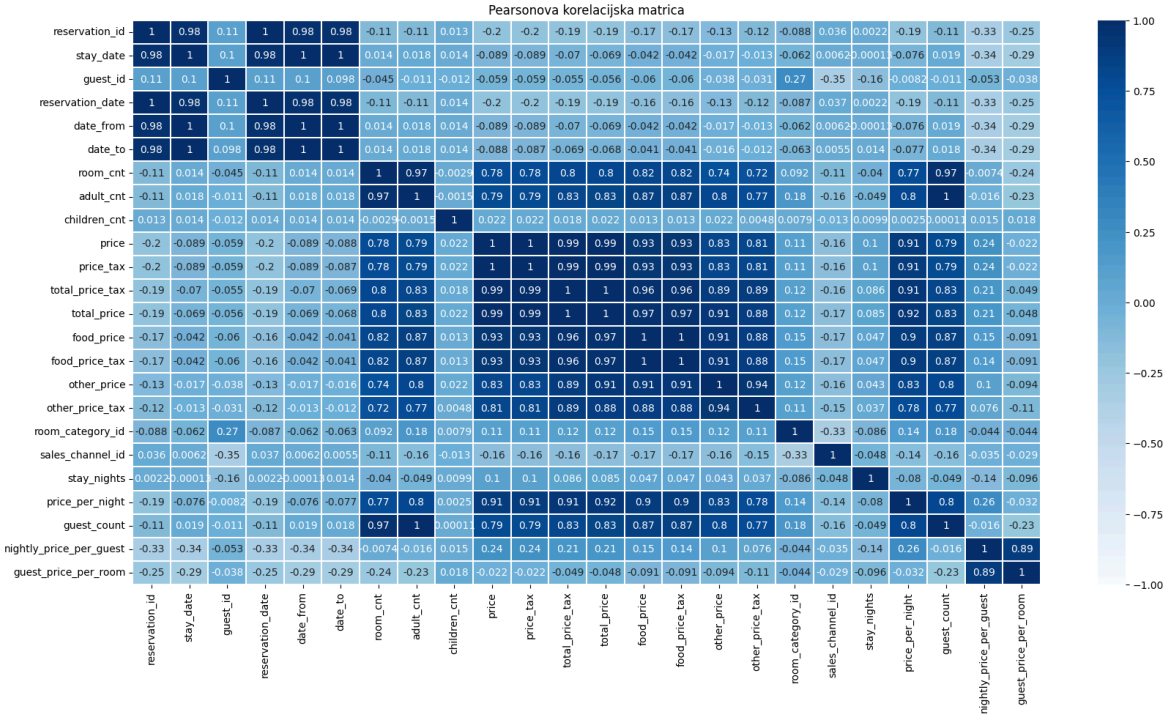

Figure 5.10.

6. Predictions

6.1. Point predictions

We have used the simple prediction model ARIMA for time series forecasting. Since the idea of this project was to learn as much as we can on statistics, predictions and the math behind it we chose to try and prepare our data manually as much as we could. We differenced the data based on days of the year (exactly one year before) and then for added stationarity we logarithmized the data. After the prediction we re-differenced the data and then manually reverted the log of data.

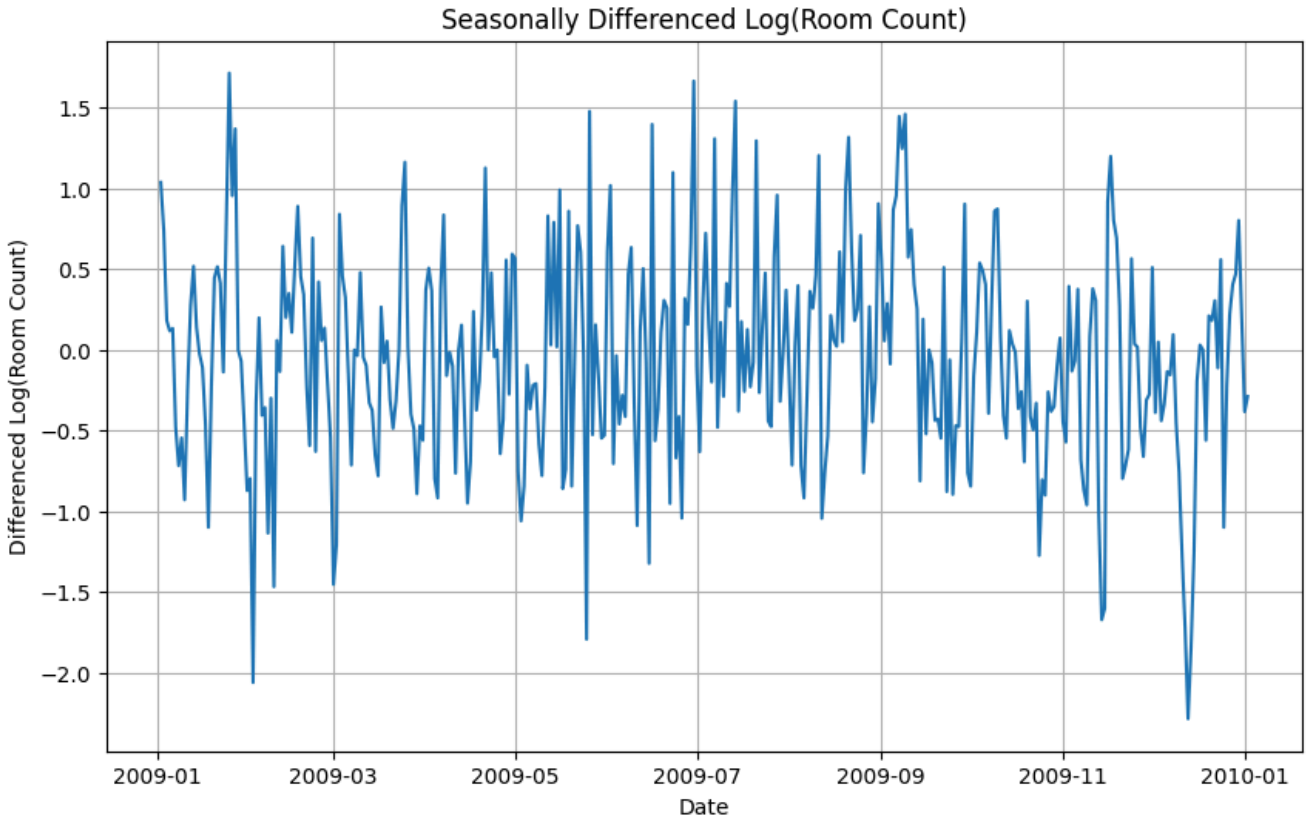

Figure 5.11.

This graph of differenced and logarithmized data has satisfactory stationarity for the ARIMA model.

After the preparation of data for our model we calculated the best parameters and trained our model. After that we got the predictions.

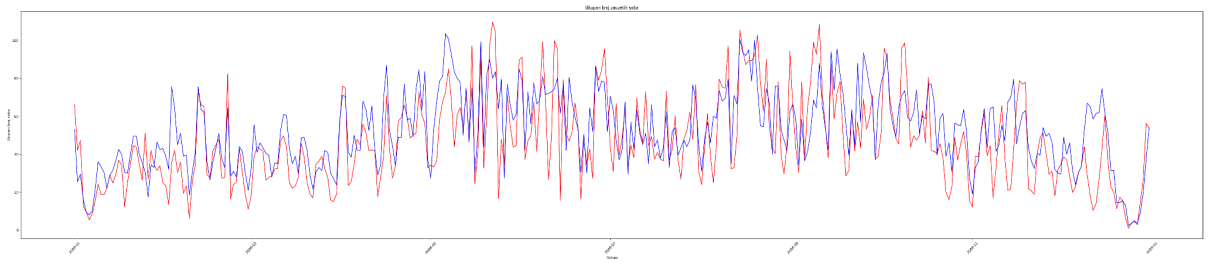

Figure 5.12.

The red line is our prediction and the blue line is our testing data set composed of mid-points of the 2 years of data we had.

6.2. Conformal prediction

We also used the mid-points for calculating the residuals and giving the conformal prediction based on residuals distribution with 95% confidence interval.

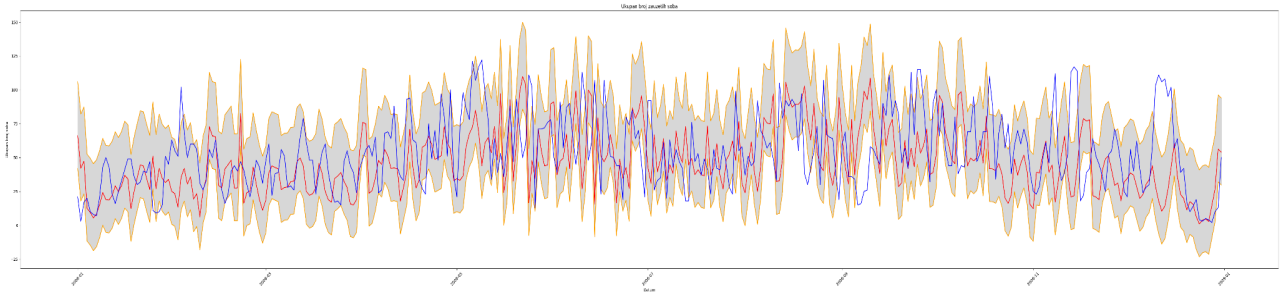

Figure 5.13.

We were satisfied with this prediction and decided to further focus on finishing the project instead of better fitting the prediction. Also better fitting doesn't mean better prediction.

7. User manual

If you want to do something similar with your specific data set you first need to be careful to name the columns the same as we had.

```
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   reservation_id      14824 non-null  int64
 1   stay_date           14824 non-null  datetime64[ns]
 2   guest_id            14824 non-null  int64
 3   guest_country_id    14824 non-null  object
 4   reservation_status  14824 non-null  object
 5   reservation_date    14824 non-null  datetime64[ns]
 6   date_from           14824 non-null  datetime64[ns]
 7   date_to             14824 non-null  datetime64[ns]
 8   resort_id           14824 non-null  int64
 9   cancel_date         2764 non-null   object
```

Figure 5.14.

Next thing would be to run the *initial_data_processing.py* for adding simple features.

After that cleaning the data is done with *data_cleanup/second_dataset/train_data_cleaning.ipynb.*

You can do your own analysis or use the data_analysis/train_data_analysis.ipynb.

And finally for the prediction you can use models/ARIMA/solution_steps.ipynb or models/ARIMA/predict.py or if you just want a prediction for our data set you can try the docker app as explained in technical documentation and README file.

8. Conclusion

We have learned a lot from this project and wish to improve on our knowledge and this project in the future. We tried every part of this project manually as much as we could and that is why we chose a simple prediction model and simple conformal prediction for better understanding of all the math and science behind it. We are satisfied with the results of the prediction and the outcome of this project.

All the contributors to this project are listed in the README file if you wish to contact us.

9. References

[1] Python - https://docs.python.org/3/

[2] pandas - https://pandas.pydata.org/docs/

[3] NumPy - https://numpy.org/doc/

[4] Jupyter - https://docs.jupyter.org/en/latest/

[5] Matplotlib - https://matplotlib.org/