**Hotel Occupancy Prediction Project**

**Introduction**

In the competitive landscape of the hospitality industry, understanding and predicting hotel occupancy rates is paramount for operational efficiency, revenue management, and enhancing guest satisfaction. As a part of a student competition we have been tasked with solving this problem. We have embarked on an project to develop a predictive model capable of forecasting hotel occupancy rates with high accuracy.

**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 ultimately improving the overall guest experience.

**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 aim 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.

**Data Cleaning Process**

**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.

**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.

**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.

**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.

**Final Data Storage:**

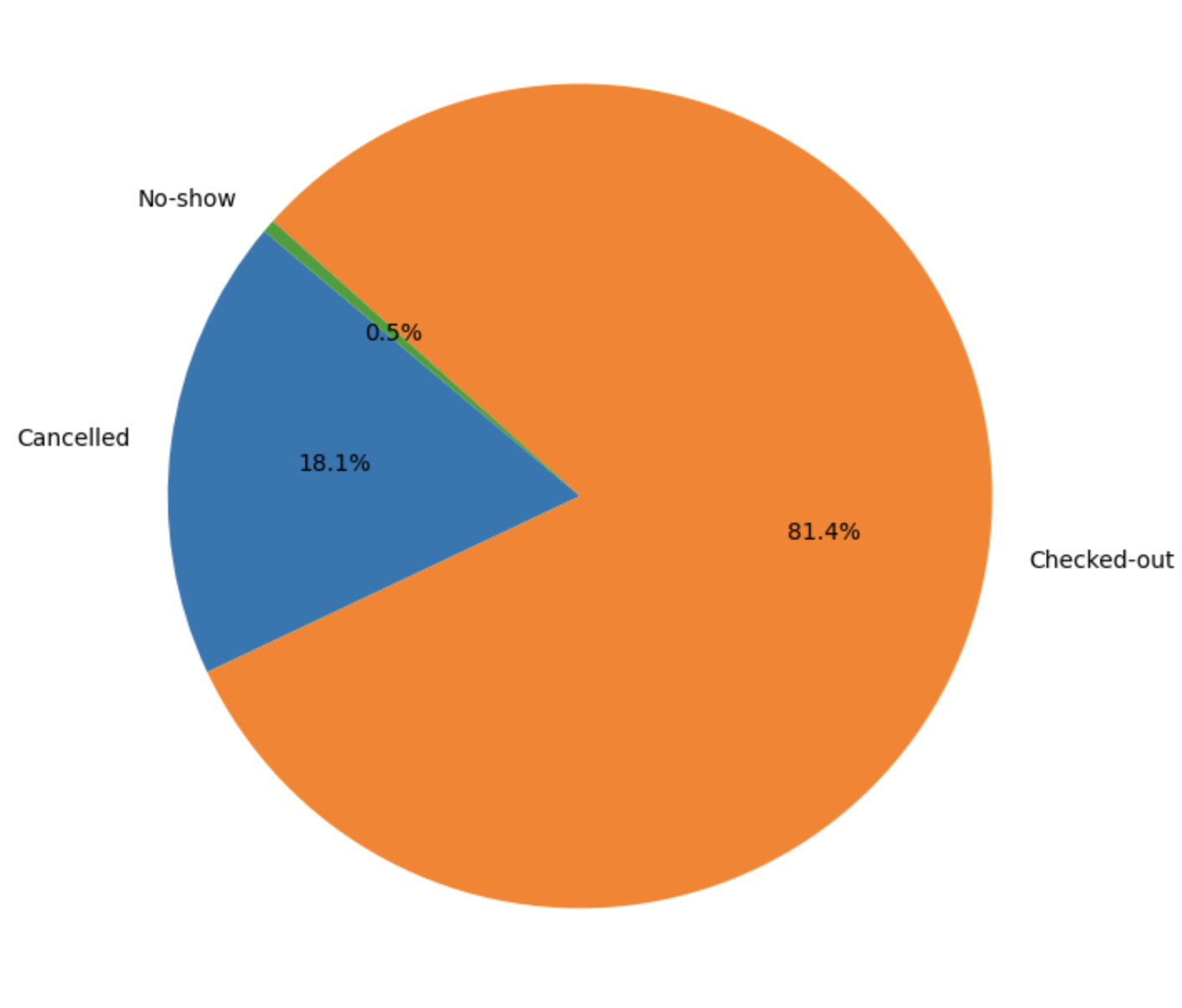
Once cleaned, the data was saved in a clean format, ready for further processing and analysis in the predictive modeling stages.

**Data Analysis**

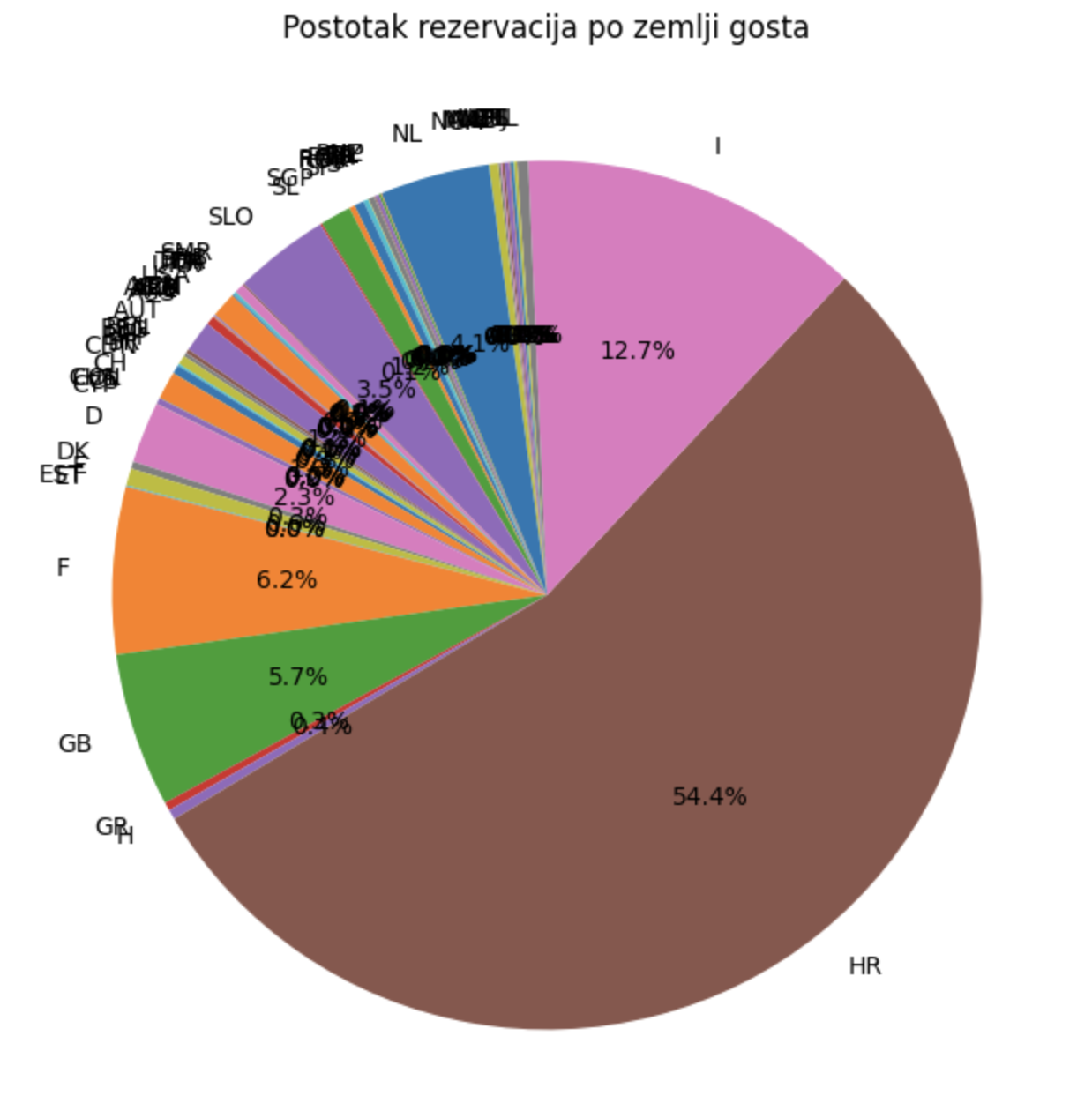
We made a comprehensive analysis of the most important variables: number of guests, county, reservation status and room type.

**General Data Overview**

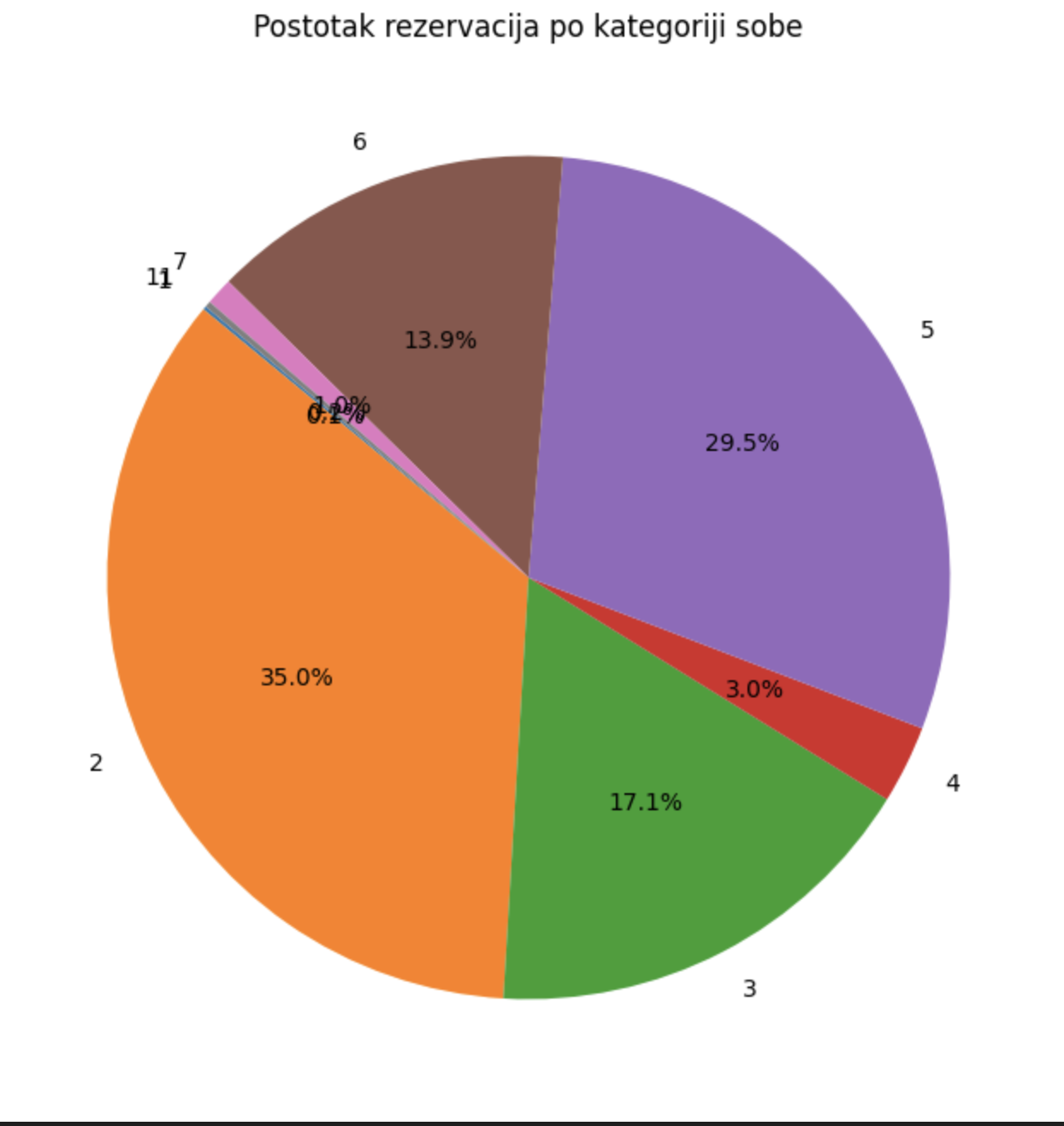
**Reservation Status**

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

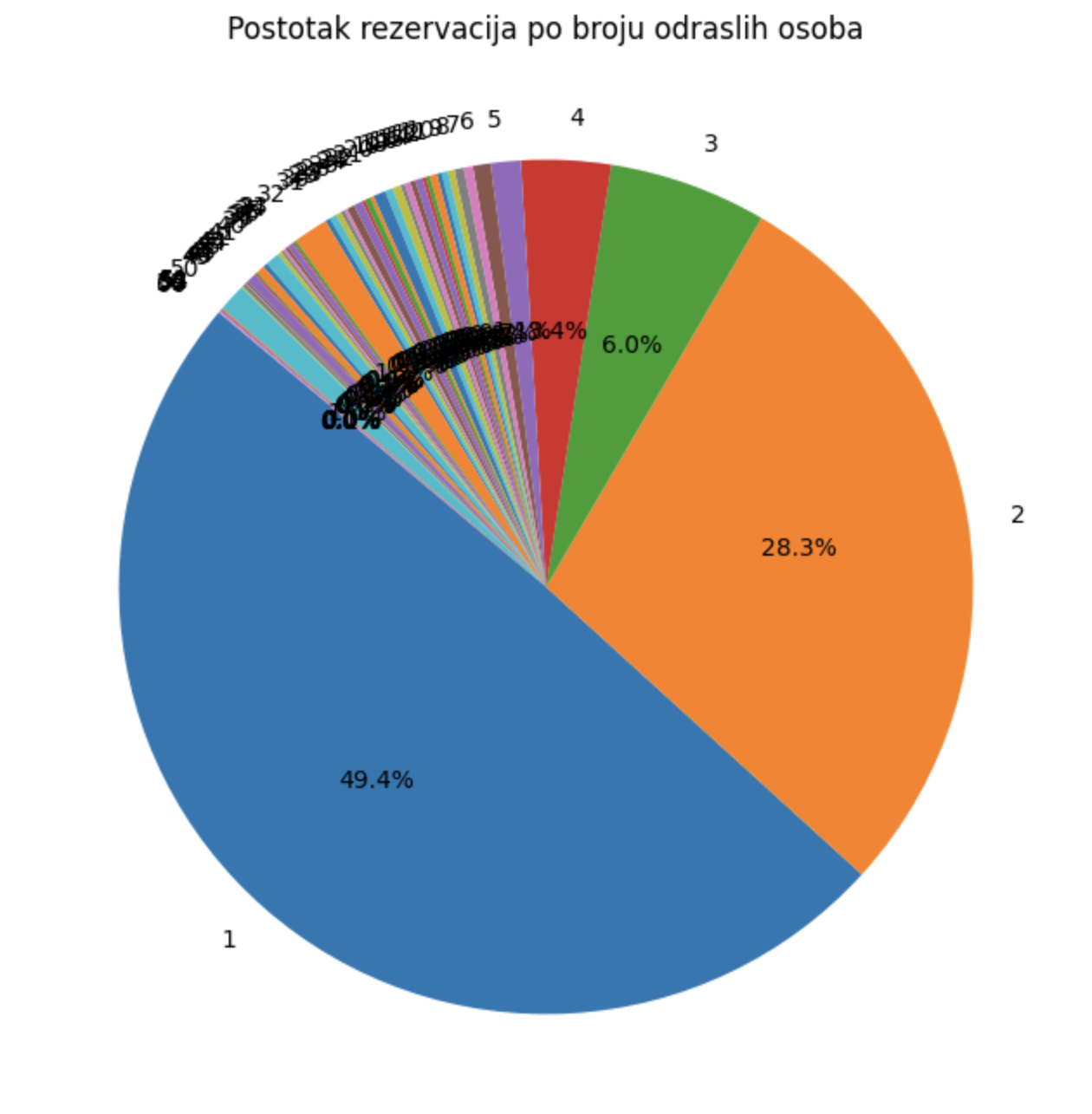
**Guest Country**

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 int he future when setting up events time table that will contain the non-working days and the holydays of the relevant countries.

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

**Room Category**: Most reservatios were made on rooms 5 and 2 indicating that this could be rooms for single person and/or cheaper rooms.

**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.

**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 singe or double bed rooms.

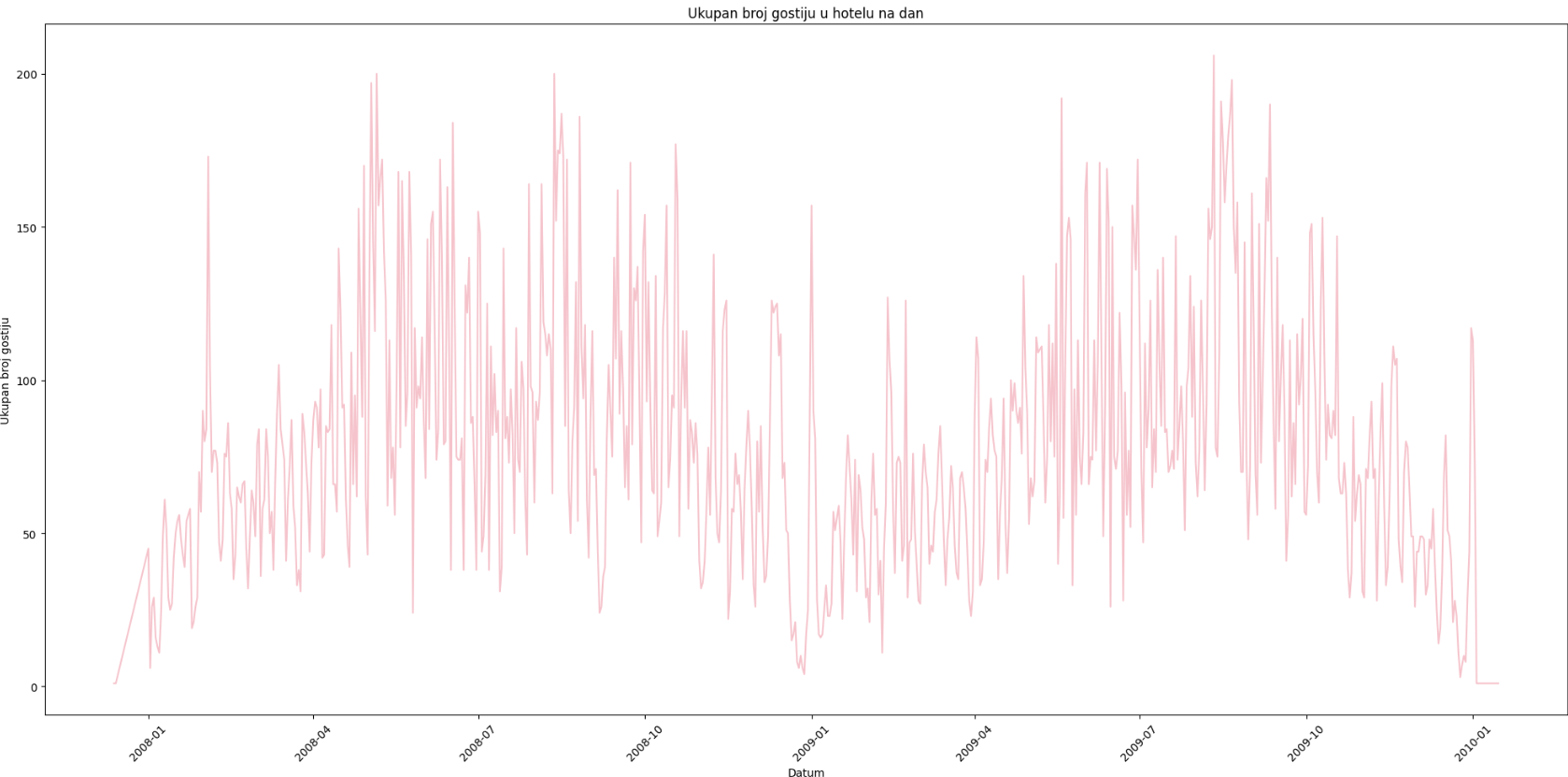
**Number of Nights Stayed**: The analysis of the stay length shows reasonable and expected patterns. Most of the reservations were made for one night (62,5%) or for 2 nights (20,1%). We found outliers in the number of nights with the maximum of over 200 days.

**\**

**Detailed Guest Data Analysis for Completed Stay**

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

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

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.