Hotel Booking Dataset
The dataset is from the article “Hotel Booking Demand Datasets” by Nuno Antonio, Ana Almeida, and Luis Nunes. The article was published in February 2019 in *Data in Brief*, Volume 22.

Context of the Dataset

The **Hotel Booking Dataset**, sourced from **Science Direct and Kaggle**, contains hotel booking data from **July 1, 2015 to August 31, 2017** for a City Hotel and a Resort Hotel. It contains **119,390 rows** and **36 columns**, providing detailed information on bookings, customer demographics, and reservation specifics.

Each record includes core details such as **hotel type** (e.g., Resort Hotel), **booking status** (canceled or not), and **lead time** (days between booking and arrival date). Temporal data covers the **year, month, week number, and day of the month for arrival**. In addition, guest-related metrics include the **number of adults, children, and babies** per booking, **meal preferences**, and whether the **guest is a returning client**. Booking and reservation features include the **booking medium** (e.g., direct or via agency), **reserved and assigned room types**, and any **booking changes.** Financial and service-specific data—such as **deposit type**, **average daily rate** (ADR), **days on the waiting list**, **parking spaces required**, and other **special requests**—add further context.

Personal data points, including **names**, **emails**, **phone numbers**, and **credit card numbers**, were also provided. However, due to privacy considerations, the original data containing customer data was removed and replaced with artificial data.

Sources:
Science Direct, https://www.sciencedirect.com/science/article/pii/S2352340918315191
Kaggle,
https://www.kaggle.com/datasets/mojtaba142/hotel-booking/data?fbclid=IwZXh0bgNhZW0CMTEAAR2aOopaxMlp5KUoYRG50Pm8cabEapQTCO04GuzPpbF-F5Mf9hnY0CWE5n0_aem_P7VwXZ9YTD726tp63iuNwA

Data Dictionary

Each observation represents a hotel booking.

Column Name	Data Type	Expected Values	Description
hotel	Text	Resort Hotel City Hotel	The hotel type
is_canceled	Number	1 0	Indicates if the booking was canceled (1) or not (0)
lead_time	Number	Non-negative Integer	Number of days between the booking date and the arrival date
arrival_date_year	Number	2015 2016 2017	Year of arrival date
arrival_date_month	Text	January February March April May June July August September October November December	Month of arrival date with 12 categories: “January” to “December”
arrival_date_week_number	Number	Non-negative Integer from 1 to 53	Week number of the arrival date in the arrival year
arrival_date_day_of_month	Number	Non-negative Integer from 1 to 31	Day of the month of the arrival date
stays_in_weekend_nights	Number	Non-negative Integer	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
stays_in_week_nights	Number	Non-negative Integer	Number of weekday nights (Monday to Friday) the guest stayed or booked to stay at the hotel
adults	Number	Non-negative Integer	Number of adults
children	Number	Non-negative Integer	Number of children
babies	Number	Non-negative Integer	Number of babies

Sources:
Science Direct, https://www.sciencedirect.com/science/article/pii/S2352340918315191
Kaggle, https://www.kaggle.com/datasets/mojtaba142/hotel-booking/data?fbclid=IwZXh0bgNhZW0CMTEAAR2aOopaxMlp5KUoYRG50Pm8cabEapQTCO04GuzPpbF-F5Mf9hnY0CWE5n0_aem_P7VwXZ9YTD726tp63iuNwA

meal	Text	Undefined SC BB HB FB	The meal package included in the booking: Undefined/SC – no meal package; BB – Bed & Breakfast; HB – Half board (breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch and dinner)
country	Text	**E.g.** PRT - Portugal GBR - United Kingdom USA - United States ESP - Spain IRL - Ireland	Country of origin of the guest/s Countries are represented in the ISO 3155– or 3166–3:2013 format
market_segment	Text	Aviation Complementary Corporate Direct Groups Offline TA/TO Online TA Undefined	The booking channel or market segment designation TA – Travel Agents; TO – Tour Operators
distribution_channel	Text	TA/TO Direct Corporate GDS Undefined	The booking distribution channel TA – Travel Agents; TO – Tour Operators; GDS – Global Distribution System
is_repeated_guest	Number	1 0	Indicates if the booking is from a repeated guest (1) or not (0)
previous_cancellations	Number	Non-negative Integer	Number of previous bookings that were canceled by the customer
previous_bookings_not_canceled	Number	Non-negative Integer	Number of previous bookings not canceled by the customer
reserved_room_type	Text	A B C D E F G H L P	Code of room type reserved (code is presented instead of designation for anonymity reasons) **Note:** Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request.

Sources:
Science Direct, https://www.sciencedirect.com/science/article/pii/S2352340918315191
Kaggle, https://www.kaggle.com/datasets/mojtaba142/hotel-booking/data?fbclid=IwZXh0bgNhZW0CMTEAAR2aOopaxMlp5KUoYRG50Pm8cabEapQTCO04GuzPpbF-F5Mf9hnY0CWE5n0_aem_P7VwXZ9YTD726tp63iuNwA

assigned_room_type	Text	A B C D E F G H I K L P	Code of room type assigned to the customer (code is presented instead of designation for anonymity reasons) **Note:** May differ from reserved type
booking_changes	Number	Non-negative Integer	Number of changes/amendments made to the booking from the moment the booking was entered on the PMS (Property Management System) until the moment of check-in or cancellation
deposit_type	Text	No Deposit Non Refund Refundable	The type of deposit made for the booking: No Deposit – no deposit was made; Non Refund – a deposit was made in the value of the total stay cost; Refundable – a deposit was made with a value under the total stay cost
agent	Number	Positive Integer	The ID of the travel agency that made the booking **Note:** An ID is presented instead of designation for anonymity reasons.
company	Number	Positive Integer	The ID of the company/entity that made or paid for the booking (if applicable) **Note:** An ID is presented instead of designation for anonymity reasons.
days_in_waiting_list	Number	Non-negative Integer	Number of days the booking was in the waiting list before it was confirmed to the customer
customer_type	Text	Contract Group Transient	The type of customer: Contract – when the booking has an allotment or other type of contract associated to it;

Sources:
Science Direct, https://www.sciencedirect.com/science/article/pii/S2352340918315191
Kaggle, https://www.kaggle.com/datasets/mojtaba142/hotel-booking/data?fbclid=IwZXh0bgNhZW0CMTEAAR2aOopaxMlp5KUoYRG50Pm8cabEapQTCO04GuzPpbF-F5Mf9hnY0CWE5n0_aem_P7VwXZ9YTD726tp63iuNwA

		Transient-party	Group – when the booking is associated to a group; Transient – when the booking is not part of a group or contract, and is not associated to other transient booking; Transient-party – when the booking is transient, but is associated to at least other transient booking
adr	Number	Non-negative float	Average Daily Rate – the average revenue per occupied room **Note:** This is calculated by dividing the sum of all lodging transactions by the total number of staying nights.
required_car_parking_spaces	Number	Non-negative Integer	Number of car parking spaces required by the customer
total_of_special_requests	Number	Non-negative Integer	Number of special requests made by the customer (e.g. twin bed or high floor)
reservation_status	Text	Canceled Check-Out No-Show	The final status of the booking: Canceled – booking was canceled by the customer; Check-Out – customer has checked in but already departed; No-Show – customer did not check-in and did inform the hotel of the reason why
reservation_status_date	Floating Timestamp	Date values in the format: YYYY-MM-DD	Date at which the last status was set
name	Text	Name of the Guest in the format: FirstName LastName	The name of the primary guest for the booking **Note:** The names are not real for anonymity reasons.
email	Text	Email address in the format: username@domainName	The email address of the guest **Note:** Username includes: ● Letters (a-z, A-Z) ● Digits (0-9) ● Special characters as long as they are not consecutive or the first or last characters: ○ Dots (.) ○ Hyphens (-)

Sources:
Science Direct, https://www.sciencedirect.com/science/article/pii/S2352340918315191
Kaggle, https://www.kaggle.com/datasets/mojtaba142/hotel-booking/data?fbclid=IwZXh0bgNhZW0CMTEAAR2aOopaxMlp5KUoYRG50Pm8cabEapQTCO04GuzPpbF-F5Mf9hnY0CWE5n0_aem_P7VwXZ9YTD726tp63iuNwA

			○ Underscores (_) Other specifics: • Must contain at least one dot (.) to separate the main domain and the top-level domain (TLD) The emails are not real for anonymity reasons.
phone-number	Text	Phone number in the Format: XXX-XXX-XXXX	The contact phone number of the guest **Note:** The phone numbers are not real for anonymity reasons.
credit_card	Text	Encrypted credit card number in the form: **********XXXX	The (masked) 16-digit credit card number associated with the booking **Note:** The credit card numbers are not real for security reasons.

Preprocessing Steps and Definitions

etl.py

EXTRACT

- Extracts the dataset from the flat file (.csv)
- Keeps a copy of the dataset through a variable df

TRANSFORM

- Transforms the dataset by applying all transformation functions in **transformations.py** to the dataset copy

transformations.py

Function 1 | **drop_columns**

Input: dataframe (df)

Returns: transformed dataframe (df)

- Drops ‘name’, ‘email’, ‘phone-number’, and ‘credit_card’ columns from the dataset

Function 2 | **change_to_string**

Input: dataframe (df)

Returns: transformed dataframe (df)

Sources:
Science Direct, https://www.sciencedirect.com/science/article/pii/S2352340918315191
Kaggle, https://www.kaggle.com/datasets/mojtaba142/hotel-booking/data?fbclid=IwZXh0bgNhZW0CMTEAAR2aOopaxMlp5KUoYRG50Pm8cabEapQTCO04GuzPpbF-F5Mf9hnY0CWE5n0_aem_P7VwXZ9YTD726tp63iuNwA

- Changes the data type of certain columns to strings to fit their categorical or descriptive nature
- The affected columns are: ['is_canceled', 'arrival_date_year', 'arrival_date_week_number', 'arrival_date_day_of_month', 'is_repeated_guest', 'agent', 'company']

Function 3 | **combine_columns**
Input: dataframe (df)
Returns: transformed dataframe (df)

- Adds an 'A–' at the beginning of each value in the 'agent' column
- Adds a 'C–' at the beginning of each value in the 'company' column
- Combines the 'agent' and 'company' columns into one column called 'booked_through'
- Replaces null values in 'booked_through' with "nan"

Function 4 | **drop_nulls**
Input: dataframe (df)
Returns: transformed dataframe (df)

- Drops all rows that contain null values from the dataset

Function 5 | **change_to_int**
Input: dataframe (df)
Returns: transformed dataframe (df)

- Changes the data type of the 'children' column into an integer as there cannot be a fraction of a child

Function 5 | **replace_undefined**
Input: dataframe (df)
Returns: transformed dataframe (df)

- Replaces all instances of the "Undefined" value from the 'meal' column into "SC" since both values represent the same thing

LOAD

- Loads the cleaned dataset into a .csv file
- Generates an output named "processed_hotel_bookings.csv"

Sources:
Science Direct, https://www.sciencedirect.com/science/article/pii/S2352340918315191
Kaggle,
https://www.kaggle.com/datasets/mojtaba142/hotel-booking/data?fbclid=IwZXh0bgNhZW0CMTEAAR2aOopaxMlp5KUoYRG50Pm8cabEapQTCO04GuzPpbF-F5Mf9hnY0CWE5n0_aem_P7VwXZ9YTD726tp63iuNwA

Output Dataset Data Dictionary

File name: processed_hotel_bookings.csv

Column Name	Data Type	Expected Values	Description
hotel	Text	Resort Hotel City Hotel	The hotel type
is_canceled	Text	1 0	Indicates if the booking was canceled (1) or not (0)
lead_time	Number	Non-negative Integer	Number of days between the booking date and the arrival date
arrival_date_year	Text	2015 2016 2017	Year of arrival date
arrival_date_month	Text	January February March April May June July August September October November December	Month of arrival date with 12 categories: “January” to “December”
arrival_date_week_number	Text	Non-negative Integer from 1 to 53	Week number of the arrival date in the arrival year
arrival_date_day_of_month	Text	Non-negative Integer from 1 to 31	Day of the month of the arrival date
stays_in_weekend_nights	Number	Non-negative Integer	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
stays_in_week_nights	Number	Non-negative Integer	Number of weekday nights (Monday to Friday) the guest stayed or booked to stay at the hotel
adults	Number	Non-negative Integer	Number of adults
children	Number	Non-negative Integer	Number of children
babies	Number	Non-negative Integer	Number of babies

Sources:
Science Direct, https://www.sciencedirect.com/science/article/pii/S2352340918315191
Kaggle, https://www.kaggle.com/datasets/mojtaba142/hotel-booking/data?fbclid=IwZXh0bgNhZW0CMTEAAR2aOopaxMlp5KUoYRG50Pm8cabEapQTCO04GuzPpbF-F5Mf9hnY0CWE5n0_aem_P7VwXZ9YTD726tp63iuNwA

meal	Text	SC BB HB FB	The meal package included in the booking: SC – no meal package; BB – Bed & Breakfast; HB – Half board (breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch and dinner)
country	Text	**E.g.** PRT - Portugal GBR - United Kingdom USA - United States ESP - Spain IRL - Ireland	Country of origin of the guest/s Countries are represented in the ISO 3155– or 3166–3:2013 format
market_segment	Text	Aviation Complementary Corporate Direct Groups Offline TA/TO Online TA Undefined	The booking channel or market segment designation TA – Travel Agents; TO – Tour Operators
distribution_channel	Text	TA/TO Direct Corporate GDS Undefined	The booking distribution channel TA – Travel Agents; TO – Tour Operators; GDS – Global Distribution System
is_repeated_guest	Text	1 0	Indicates if the booking is from a repeated guest (1) or not (0)
previous_cancellations	Number	Non-negative Integer	Number of previous bookings that were canceled by the customer
previous_bookings_not_canceled	Number	Non-negative Integer	Number of previous bookings not canceled by the customer
reserved_room_type	Text	A B C D E F G H L P	Code of room type reserved (code is presented instead of designation for anonymity reasons) **Note:** Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request.

Sources:
Science Direct, https://www.sciencedirect.com/science/article/pii/S2352340918315191
Kaggle, https://www.kaggle.com/datasets/mojtaba142/hotel-booking/data?fbclid=IwZXh0bgNhZW0CMTEAAR2aOopaxMlp5KUoYRG50Pm8cabEapQTCO04GuzPpbF-F5Mf9hnY0CWE5n0_aem_P7VwXZ9YTD726tp63iuNwA

assigned_room_type	Text	A B C D E F G H I K L P	Code of room type assigned to the customer (code is presented instead of designation for anonymity reasons) **Note:** May differ from reserved type
booking_changes	Number	Non-negative Integer	Number of changes/amendments made to the booking from the moment the booking was entered on the PMS (Property Management System) until the moment of check-in or cancellation
deposit_type	Text	No Deposit Non Refund Refundable	The type of deposit made for the booking: No Deposit – no deposit was made; Non Refund – a deposit was made in the value of the total stay cost; Refundable – a deposit was made with a value under the total stay cost
booked_through	Text	A–(Non-negative Integer) **E.g.** **A-110** **A-328** C–(Non-negative Integer) **E.g.** **C-110** **C-328**	The ID of the travel agency or company that made the booking If the booking was made by a travel agent, its ID is prefixed by "A–". If the booking is done by a company, its ID is prefixed by "C–". **Note:** An ID is presented instead of designation for anonymity reasons.
days_in_waiting_list	Number	Non-negative Integer	Number of days the booking was in the waiting list before it was confirmed to the customer
customer_type	Text	Contract Group Transient	The type of customer: Contract – when the booking has an allotment or other type of contract associated to it;

Sources:
Science Direct, https://www.sciencedirect.com/science/article/pii/S2352340918315191
Kaggle, https://www.kaggle.com/datasets/mojtaba142/hotel-booking/data?fbclid=IwZXh0bgNhZW0CMTEAAR2aOopaxMlp5KUoYRG50Pm8cabEapQTCO04GuzPpbF-F5Mf9hnY0CWE5n0_aem_P7VwXZ9YTD726tp63iuNwA

		Transient-party	Group – when the booking is associated to a group; Transient – when the booking is not part of a group or contract, and is not associated to other transient booking; Transient-party – when the booking is transient, but is associated to at least other transient booking
adr	Number	Non-negative float	Average Daily Rate – the average revenue per occupied room **Note:** This is calculated by dividing the sum of all lodging transactions by the total number of staying nights.
required_car_parking_spaces	Number	Non-negative Integer	Number of car parking spaces required by the customer
total_of_special_requests	Number	Non-negative Integer	Number of special requests made by the customer (e.g. twin bed or high floor)
reservation_status	Text	Canceled Check-Out No-Show	The final status of the booking: Canceled – booking was canceled by the customer; Check-Out – customer has checked in but already departed; No-Show – customer did not check-in and did inform the hotel of the reason why
reservation_status_date	Floating Timestamp	Date values in the format: YYYY-MM-DD	Date at which the last status was set

Sources:
Science Direct, https://www.sciencedirect.com/science/article/pii/S2352340918315191
Kaggle, https://www.kaggle.com/datasets/mojtaba142/hotel-booking/data?fbclid=IwZXh0bgNhZW0CMTEAAR2aOopaxMlp5KUoYRG50Pm8cabEapQTCO04GuzPpbF-F5Mf9hnY0CWE5n0_aem_P7VwXZ9YTD726tp63iuNwA