FINAL PROJECT – DELIVERABLE 1

Dhruvi Chauhan, Genglin Liu, Vinitha Maheswaran, Anushka Neerav Doshi

Description of the dataset (5%)

We will be working with the **'Hotel Booking Demand Dataset'**. It is a publicly available dataset (can be found on Kaggle) containing booking transactions from a city hotel and a resort hotel in Portugal. The dataset contains 119390 observations and 32 attributes including information such as when the booking was made, length of stay, the number of adults, children, and/or babies, the number of available parking spaces, and many others for the respective hotels. Each observation represents a hotel booking. Since this is a public dataset, all data elements pertaining to hotel or customer identification were deleted. This dataset can have an important role for research and educational purposes in revenue management, machine learning, data mining and other fields.

Users identification (10%)

1) People who are booking hotels

Background: Tourists can have varying knowledge bases depending on whether they are locals or foreigners, visited the place before, etc. We don't expect the general population to be particularly knowledgeable about analyzing the dataset, they would rather just need to capture the key information from the presentation of the data (price, location etc)
Motivations: Cost (best experience at optimal price), amenities, location, room choices, check-in/out times, lead time, reviews, car parking spaces, cancellation policy (refund/deposit), meal options, safety, agencies for booking, days in waitlist before getting confirmed, total number of special requests, reserved room same as assigned room
Tasks: To check and understand trends of hotel price over a period of time and plan their travel accordingly within budget; learn about hotel accommodations and features before booking, etc.
Context of use: When they go on a website to book a hotel or when they are planning an upcoming trip.
Challenges: Complex visualization, while containing a richer amount of information, could have adversarial effects and be more confusing for the general public who are not experts in reading visualizations. Some users might not have much experience with different visualization types.

2) Hotel management

Background: Hotel management is the people who manage a specific hotel. Hotel management team may include the executives, front desk services, housekeeping and maintenance, accounting, marketing and sales, and kitchen staff. They are knowledgeable about the local hotel services and their local competitors. They are particularly interested in the prices and services that they (the given particular hotel) and the nearby hotels (local competitors) can provide.

Motivations: To understand what motivates people to choose a certain hotel, trends in booking, popular seasons / weekdays or weekends (to set prices), to improve facilities they can provide (Ex: shuttle service), why do people visit (vacation, conferences etc), understanding when do people come based on the purpose of their visit (for vacation, holiday season; for business meetings, other times of the years), cancellations, total number of special requests
Tasks: To keep track of the relevant information about themselves as well as their competitors. Understand and analyze the seasonal trend of hotel booking and accommodate different types of visitors that they have (e.g. provide multilingual support for foreign visitors)
Context of use: Hotel management could use such visualization in their monthly report among the staff and use this information to determine room prices and adjust their service / marketing strategy.
Challenges: From the given dataset the hotel management might not be able to obtain too much information about their competitors. Nonetheless the hotel management could always learn useful information from the existing data across a long period of time, and use that for their own sales strategy.

3) Tourist / hospitality services - (e.g. travel agency, airlines / car rental companies)

Background: These users are the people who are very familiar with the local hotels and they might have collaborations with multiple hotels. They are interested in how the local hotels are doing collectively and may be less interested in the behavior of a single hotel. If the hotels in a given area are having increasing bookings at a time, then they would react accordingly.
Motivations: popular seasons/weekdays or weekends, check-in/out times, trends in agents chosen (competitors), waitlist, target customers
Tasks: To observe the times when hotels in the region are in high demand, analyze the duration of typical stays, and use the information to help plan their own services.
Context of use: Accessing publicly available records on hotel data in the region and using it to aid their own sales plan. Using these reports to make assessments about the popular seasons and target customers during their own discussion.
Challenges: Publicly available data may not always be sufficient for hospitality services; in order to make more informed decisions, they will likely need to obtain more data than what is available in this dataset.

4) Local Government / independent data analysts

Background: Local government or independent data analysts are the people who are skilled to work with the dataset but not specific to the hotel data. The public sector would be interested in regulating or overseeing the local economy and overall tourism development and hospitality service industry. Independent analysts are interested in finding the broad trends of the data but not specific to any particular travel planning detail or an individual hotel.
Motivations: ratio of local/foreign visitors (which country do most tourists come from), average daily rates - revenue, tourist season, overseeing the distribution of the hotels in an area,

proportions of different market segments, demands (seasonality) of hotel services over a long period of time.
Tasks: Observing the overall trend of tourist activities in the region and analyzing the different types of visitors in the hotels during different seasons.
Context of use: The branch of the local government that oversees hospitality services might be using the hotel data for their own seasonal/quarterly reports. Independent analysts may be monitoring this data regularly and extracting useful information for their own specified tasks.
Challenges: A similar problem is that in order to form more accurate analysis on a broader scale, they would need to have more data that contains information about more hotels in a region.

Task identification (10%)

People who are booking hotels:

1. People looking to stay at a hotel would want to know an optimal time to make reservations for. This data would help them make a decision. They can also gauge the most popular seasons for the different hotels.
2. Visitors would want to observe the trends in the average daily cost of different rooms at different hotels in order to make a choice with good value for money.
3. Visitors would want their chosen hotel to have the required amenities or facilities. This data would allow them to explore the different rooms available and options for meals, parking spaces provided, special requests met, etc. and perhaps correlate that with the cost.

Hotel management:

- The hotel revenue manager wants to discover the trends in bookings to identify the peak/holiday seasons or whether more rooms are booked for the weekdays/weekends and determine the price of hotel rooms accordingly.
- The hotel front desk wants to predict whether they are likely to receive a disproportionately high number of special requests depending on the customer type and the number of kids or children who will be staying. (Satisfying special requests would ensure the hotel management that the customers will have an enjoyable stay.)
- The hotel management would want to identify the purpose of their customers' visit and improve facilities and provide additional services to ensure that the prospects are converted into paying customers.
- The hotel revenue manager/executive would want to discover the trends in room cancellation and determine the deposit amount to be charged for the rooms depending on the season and time of the week.

Tourist/hospitality services:

1. Travel agencies want to compare their standing in the market as opposed to their competitors and study what possible factors might lead to a customer selecting another agency over them.
2. Travel agents and/or Tour operators want to discover trends in hotel booking rates over different months of the year so they can adjust their customer pricing given a season in order to increase profit. They also want to locate hotels that are more popular to create ties with them.
3. Tour operators want to explore the preference of hotel type (resort vs. city hotel) by a specific customer_type and their place of origin so as to design itineraries and packages that they can sell to a customer or a travel agency.

Local Government / independent data analysts:

1. Local governments would be interested in monitoring or overseeing the local economy and overall tourism development and hospitality service industry. Specifically, the local government might be interested in the regional distribution of hotels, such as where hotels are clustered and which region has more (or less) expensive hotels.
2. The department that oversees hospitality would probably like to learn about the seasonal revenue that a region makes collectively from hospitality services and tourism, and that information could be reflected in the hotel data.
3. Similar to the interests of the local government, independent analysts are interested in finding the broad trends of the data and may not be concerned with the specific performance of an individual hotel.
4. Analysts might also be interested in making inferences beyond the hotels, such as what kind of travelers do hotels in a given region usually have, are there a lot of children in those hotels, are a lot of the travelers from another country, etc.

Data Abstraction

1. The dataset type is a table because it only has attributes and Items.
2. There are 32 fields/attributes.
3. There are 119390 distinct rows and thus 119390 items.

Attribute Description:

- **hotel** : Hotel (Resort Hotel or City Hotel)
    - Type: Categorical
    - Cardinality: 2 (unique values)
- **is_canceled** : Value indicating if the booking was canceled (1) or not (0)
    - Type: Categorical (binary)
    - Cardinality: 2 (unique values)

- **lead_time** : Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
    - Type: Quantitative (Sequential)
    - Cardinality: 479 (unique values)
    - Range: 0 - 737
- **arrival_date_year** : Year of arrival date
    - Type: Ordinal (Sequential)
    - Cardinality: 3 (unique values)
    - Range: 2015 - 2017
- **arrival_date_month** : Month of arrival date
    - Type: Ordinal (Cyclic)
    - Cardinality: 12 (unique values)
- **arrival_date_week_number** : Week number of year for arrival date
    - Type: Ordinal (Cyclic)
    - Cardinality: 53 (unique values)
    - Range: 1 - 53
- **arrival_date_day_of_month** : Day of arrival date
    - Type: Ordinal (Cyclic)
    - Cardinality: 31 (unique values)
    - Range: 1 - 31
- **stays_in_weekend_nights** : Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
    - Type: Ordinal
    - Cardinality: 17 (unique values)
    - Range: 0 - 19
- **stays_in_week_nights** : Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
    - Type: Ordinal
    - Cardinality: 35 (unique values)
    - Range: 0 - 50
- **adults** : Number of adults
    - Type: Ordinal
    - Cardinality: 14 (unique values)
    - Range: 0 - 55
- **children** : Number of children
    - Type: Ordinal
    - Cardinality: 5 (unique values)
    - Range: 0 - 10
    - Notes: 4 N/A or missing values present in this attribute.
- **babies** : Number of babies
    - Type: Ordinal
    - Cardinality: 5 (unique values)
    - Range: 0 - 10

- **meal** : Type of meal booked. Categories are presented in standard hospitality meal packages:
    - 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)
        - Type: Categorical
        - Cardinality: 5 (unique values)
- **country** : Country of origin. Categories are represented in the ISO 3155–3:2013 format
    - Type: Categorical
    - Cardinality: 177 (unique values)
    - Notes: 488 missing values present in this attribute
- **market_segment** : Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”
    - Type: Categorical
    - Cardinality: 8
- **distribution_channel** : Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”
    - Type: Categorical
    - Cardinality: 5
- **is_repeated_guest** : Value indicating if the booking name was from a repeated guest (1) or not (0)
    - Type: Categorical (binary)
    - Cardinality: 2
- **previous_cancellations** : Number of previous bookings that were canceled by the customer prior to the current booking
    - Type: Ordinal
    - Cardinality: 15 (unique values)
    - Range: 0 - 26
- **previous_bookings_not_canceled** : Number of previous bookings not canceled by the customer prior to the current booking
    - Type: Ordinal
    - Cardinality: 73 (unique values)
    - Range: 0 - 72
- **reserved_room_type** : Code of room type reserved. Code is presented instead of designation for anonymity reasons.
    - Type: Categorical
    - Cardinality: 10 (unique values)
- **assigned_room_type** : Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons.
    - Type: Categorical

    - Cardinality: 12 (unique values)
- **booking_changes** : Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
    - Type: Ordinal
    - Cardinality: 21 (unique values)
    - Range: 0 - 21
- **deposit_type** : Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories:
    - 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 cost of stay.
        - Type: Categorical
        - Cardinality: 3 (unique values)
- **agent** : ID of the travel agency that made the booking
    - Type: Categorical
    - Cardinality: 333 (unique values)
    - Notes: 16340 NULL / missing values in this attribute.
- **company** : ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons
    - Type: Categorical
    - Cardinality: 352 (unique values)
    - Notes: 112593 NULL / missing values in this attribute.
- **days_in_waiting_list** : Number of days the booking was in the waiting list before it was confirmed to the customer
    - Type: Quantitative (Sequential)
    - Cardinality: 128 (unique values)
    - Range: 0 - 391
- **customer_type** : Type of booking, assuming one of four categories:
    - Contract - when the booking has an allotment or other type of contract associated to it
    - 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
        - Type: Categorical
        - Cardinality: 4 (unique values)
- **adr** : Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
    - Type: Quantitative
    - Cardinality: 8879 (unique values)
    - Range: (-6.38) to 5400

- **required_car_parking_spaces** : Number of car parking spaces required by the customer
    - Type: Ordinal
    - Cardinality: 5 (unique values)
    - Range: 0 - 8
- **total_of_special_requests** : Number of special requests made by the customer (e.g. twin bed or high floor)
    - Type: Ordinal
    - Cardinality: 6 (unique values)
    - Range: 0 - 5
- **reservation_status** : Reservation last status, assuming one of three categories:
    - 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
        - Type: Categorical
        - Cardinality: 3 (unique values)
- **reservation_status_date** : Date at which the last status was set. This variable can be used in conjunction with the *ReservationStatus* to understand when was the booking canceled or when did the customer checked-out of the hotel
    - Type: Ordinal (Sequential)
    - Cardinality: 926 (unique values)
    - Range: 10/17/2014 to 09/14/2017

Task Abstraction

People who are booking hotels:

1. What is the average daily cost for the different rooms in each hotel?
    - **Chart:** A chart will be useful to answer this question, perhaps a grouped bar chart, since it can show the change in the rates for the different rooms in each hotel and can be compared easily
    - **Attributes:** The attributes that would be required are: “hotel”, “assigned room type” and “adr”.
    - **Transformations:** The data would have to be grouped by hotel. The attribute and dataset types will remain the same
    - All the data needed to answer this question is available.
    - **Munzner’s Taxonomy**: explore trends, compare distribution

2. What facilities are available or met at each hotel?
    - **Chart:** Since there are many dimensions, a chart will be required to answer this question, like a map or interactive chart
    - **Attributes:** The attributes that would be required are: “hotel”, “meal” “required_car_parking_spaces” and “total_of_special_requests”

- All the data needed to answer this question is available.
- **Transformations:** We would have to group by hotel, and aggregate over each attribute representing the facility. The attribute and dataset types will remain the same
- **Munzner's Taxonomy**: compare distribution, locate outliers

3. For each hotel, when is the most popular time to visit in a week, month and year?
    - **Chart:** Since there are many dimensions, a chart will be required to answer this question. A line chart will be useful to observe trends
    - **Attributes:** The attributes that would be required are: "hotel", "arrival_date_month", "arrival_date_year", "arrival_date_day_of_month", "arrival_date_week_number"
    - **Transformations:** We would have to group by hotel. The attribute and dataset types will remain the same
    - All the data needed to answer this question is available.
    - **Munzner's Taxonomy**: discover trends, compare distribution

Hotel management:

1. What is the weekly, monthly, quarterly and annual demand for rooms for the two hotels? Are more bookings recorded during the weekdays or weekends? How should the hotel rooms be priced according to the demand?
    - **Chart:** We would need charts (line charts or bar charts) to answer these questions. A line chart will be helpful in analyzing the trends and identifying the peaks. During the peaks and troughs, the hotel revenue management can accordingly increase or decrease the room prices. Bar charts can be used to compare the weekly demand (whether weekdays or weekends have more bookings.)
    - **Attributes:** "hotel", "arrival_date_month", "arrival_date_year", "arrival_date_day_of_month", "arrival_date_week_number", "stays_in_weekend_nights", "stays_in_week_nights" and "adr"
    - **Transformations:** We need to group by hotel followed by (i) group by arrival date week number to get count of weekly bookings (ii) group by arrival date month to get count of monthly bookings (iii) group by arrival date month by groups of 3 months to get count of quarterly bookings (iv) group by arrival date year to get count of yearly bookings.
    - The attribute type will remain the same and so will the dataset type.
    - All the data needed to answer this question is available.
    - **Munzner's Taxonomy:** Discover trends, Compare distributions, Locate outliers, Identify value,

2. What is the purpose of the visit of the customers? What is the distribution of the customer types? Depending on the customer type, are they traveling with babies and/or children and do they require special requests and additional services?

- **Chart:** An interactive chart will most likely be needed to answer this question since it has many dimensions.
- **Attributes:** “hotel”, “company”, “customer_type”, “total_of_special_requests”, “required_car_parking_spaces”, “adults”, “children”, “babies”, “meal” are the attributes required. The purpose of the visit of the customers can be identified using whether the booking was made by a company (which indicates it is a business trip) or whether the adults are accompanied by children/babies (which indicate it is a vacation). The hotel may expect a higher number of special requests if the number of babies/children among the guests are high. Guests who came for business purposes may expect services like car parking spaces and meals.
- **Transformations:** We would need to group by hotel and company and aggregate and count the observations with values for the additional attributes mentioned above.
- The attribute type will remain the same and so will the dataset type.
- All the data needed to answer this question is available.
- **Munzner’s Taxonomy**: Discover distribution and Explore dependency.

3. What is the distribution of cancellations for the hotels? Is there any relationship between lead time and cancellations or between deposit type and cancellations? Should the hotel management charge any cancellation charges to prevent frequent cancellations by the repeated customers?
    - **Chart:** We would need charts to answer these questions. Scatterplots can be used to find any correlations between lead time and cancellations or between deposit type and cancellations. Bar charts can be used to see the distribution of deposit type.
    - **Attributes:** “hotel”, “is_canceled”, “lead_time”, “is_repeated_guest”, “previous_cancellations”, “deposit_type” are the required attributes.
    - **Transformations:** We would need to group by hotel and for each hotel count the observations based on deposit type.
    - The attribute type will remain the same and so will the dataset type.
    - All the data needed to answer this question is available.
    - **Munzner’s Taxonomy:** Discover correlations, Identify features

Tourist and hospitality services - Market Segments

1. What market segment (Online/Offline TA, TO, etc.) is contributing the most in the hotel booking process? Which segments usually have more cancellations and how do the ADR’s compare across the different segments?
    - **Chart:** A bar chart or histogram will be best to answer this question
    - **Attributes:** The attributes that we explore here are: “hotel”, “market_segment”, “adr”, and “is_canceled”.

- **Transformations:** We would need to segregate the hotels and for each hotel aggregate and count the rows grouped by market segment and cancellations and also average the adr for each segment.
- The attribute type will remain the same and so will the dataset type. However, we will introduce new variables like % demand of hotel given market segment.
- All the data needed to answer this question is available.
- **Munzner's Taxonomy**: explore and identify trends, compare distribution.

2. What is the weekly, monthly, yearly seasonal demand of the more popular market segments of the two hotels?
    - **Chart:** A chart will be needed to answer this question since it has many dimensions. A line chart will be helpful in seeing the trends and the rise and drop of popularity of the hotels. A bar chart can show the weekly demand.
    - **Attributes:** "hotel", "market_segment", "arrival_date_month", "arrival_date_year", "arrival_date_day_of_month", "arrival_date_week_number"
    - **Transformations:** There will be grouping of hotel and segment and then aggregation of some sort on the rest of the attributes depending on the type of visualization.
    - The attribute type will remain the same and so will the dataset type.
    - All the data needed to answer this question is available.
    - **Munzner's Taxonomy**: study trends and locate outliers.

3. **[Tour Operators]** What is the distribution of the customer types across the two hotels and where do these customers come from? Depending on the customer type, are they traveling with babies and/or children? If so, do they need extra amenities?
    - **Chart:** An interactive chart will most likely be needed to answer this question since it has many dimensions.
    - **Attributes:** The attributes that we explore here are: "hotel", "customer_type", and "country". Additionally, "required_car_parking_spaces", "total_of_special_requests", "reserved_room_type", "adults", "children", "babies", "meal" are attributes that can be used to add more layers and depth to the question.
    - **Transformations:** We would need to aggregate and count the rows with the same values for the additional attributes described above grouped according to the hotel and the country of origin as well as the customer type.
    - The attribute type will remain the same and so will the dataset type.
    - All the data needed to answer this question is available.
    - **Munzner's Taxonomy**: discover distribution and explore dependency.

Local Government / independent data analysts:

1. What are the countries that most hotel customers come from? How to rank the countries in the order of their visitors in the resort and city hotel?
    - **Chart**: We could need a chart to answer this question, e.g. a pie chart or a map.

- **Data transformation**: we might need to extract out a specific column of data, and rank or sort them in descending order.
- **Attribute**: For this question we would be particularly interested in the “country” attribute. And we would also like to have the “hotel type” so we can group by types of the hotel.
- The data attribute type would not change when we transform it
- We have all the data we need for this question
- **Munzner’s Taxonomy**: identify value, compare distribution

2. What is the seasonal revenue generated by the resort hotel vs city hotel? Could we observe a trend of seasonality and predict how busy the hotel is in a given month or week of the year?
    - **Chart**: We could use a chart to visualize the revenue query, though in theory the values could be computed. For the trend, a visualization is needed, e.g. a line chart or histogram.
    - **Data transformation:** we would group the data by seasons and calculate the accumulated revenue from both hotels, then present them in the order of months in a year.
    - **Attribute**: need the “ADR”, “arrival_date_year”, “arrival_date_month”, “arrival_date_week”
    - The attribute types don’t change when we transform it
    - We have all the data we need for this question
    - **Munzner’s Taxonomy**: discover trend, compare value

3. What does the hotel data tell about the local economy?
    - **Chart**: we should use a chart to visualize the information we gather. This is a more open-ended question and asks for more exploratory analysis.
    - **Data transformation:** we would group the data by their types and calculate the accumulated revenue from both hotels, as well as the composition of their visitors (number of foreigners, adults and children), plus the booking information by market segment.
    - **Attribute**: need the “ADR”, “market_segment”, “adults”, “children”, “babies”, “reserved_room_type”. From these attributes we can gather information about the hotel revenues, how much they interact with other market segments, and what type of room (more expensive or less expensive) people book there, the number of solo travelers vs families, and make inferences about how this information would reflect the local economy.
    - The attribute types don’t change when we transform it
    - We have most of the data we need for this question, although since this is more open-ended, we could always use more attributes even if they are not present in this dataset, such as information about more hotels in the region.
    - **Munzner’s Taxonomy**: discover distribution, compare value, derive feature, explore dependency