**Capstone Project Technical Documentation:**

|  |
| --- |
| **Name, Email and Contribution:** |
| 1. Abhishek Kumar Mishra ([Abhishekkumarmishra364@gmail.com](mailto:Abhishekkumarmishra364@gmail.com))  * Data Wrangling * Data Mining * Data Analysis * Visualizations * Observations * Summarization * Conclusions |
| **GitHub Repo link.** |
| Github Link:- https://github.com/Abhishek709mis/Hotel-Booking-Analysis  Drive Link:- |
| * **Writing a short summary of your Capstone project and its components.** * **Describe the   a) problem statement,   b) approaches and**   **c) Conclusions.** |
| The Customer’s expectation of a hotel experience has grown over the years and hotel managements need to keep up with it. As the per hotel is good used in bussniess and have a good hotel as you want the AI and data science provide. The main objective is explore the data and explore the insight knowledge, so that which boost the hostel bussniness.  “Travel is my therapy..”  From the quotes I can say that Travelling is habbit to humans. People wish to connect, explore more. When there is not the data science than people does not know how to attract and get more number of customer and why people prefer than my hotel.Data science-play a major role. Nowadays, having a proper understanding of data and connecting all data sources effectively is paramount in generating competitive advantage.  **Importing Libraries:**  Some of the libraries like Num-Py for numerical operations, Pandas for data manipulation, matplotlib and seaborn for data visualization were loaded. In addition to these libraries, pycountry library was installed and loaded.  **Reading Data:**  After drive was mounted, data from csv file was read and store in a pandas dataframe.  **Data Inspection:**  Loading the dataset and importing relevant libraries, dataset has been explored by thoroughly by looking into its head, tail, brief summary, number of records and features, etc. The data-frame contains 119390 rows of data, out of which 31,994 rows are duplicate rows, which must be removed later. The data-frame contains 32 columns. Four columns have missing values. Some columns require conversion of data-types.  The features in the dataset were identified as:   * 1. hotel: Name of the hotel (Resort Hotel or City Hotel).   2. is\_canceled: If the booking was canceled (1) or not (0).   3. lead\_time: Number of days before the actual arrival of the guests.   4. arrival\_date\_year: Year of arrival date.   5. arrival\_date\_month: Month of arrival date.   6. arrival\_date\_week\_number: Week number of year for arrival date.   7. arrival\_date\_day\_of\_month: Day of month arrival date.   8. stays\_in\_weekend\_nights: Number of weekend nights (Saturday or Sunday) spent at the hotel by the guests.   9. stays\_in\_week\_nights: Number of weeknights (Monday to Friday) spent at the hotel by the guests.   10. adults: Number of adults among guests.   11. children: Number of children among guests.   12. babies: Number of babies among guests.   13. meal: Type of meal booked.   14. country: Country of guests.   15. market\_segment: Designation of market segment.   16. distribution\_channel: Name of booking distribution channel.   17. is\_repeated\_guest: If the booking was from a repeated guest (1) or not (0).   18. previous\_cancellations: Number of previous bookings that were cancelled by the customer prior to the current booking.   19. previous\_bookings\_not\_canceled: Number of previous bookings not cancelled by the customer prior to the current booking.   20. reserved\_room\_type: Code of room type reserved.   21. assigned\_room\_type: Code of room type assigned.   22. booking\_changes: Number of changes/amendments made to the booking.   23. deposit\_type: Type of the deposit made by the guest.   24. agent: ID of travel agent who made the booking.   25. company: ID of the company that made the booking.   26. days\_in\_waiting\_list: Number of days the booking was in the waiting list.   27. customer\_type: Type of customer, assuming one of four categories.   28. adr: Average Daily Rate, as defined by dividing the sum of all lodging transactions by the total number of staying nights.   29. required\_car\_parking\_spaces: Number of car parking spaces required by the customer.   30. total\_of\_special\_requests: Number of special requests made by the customer.   31. reservation\_status: Reservation status (Canceled, Check-Out or No-Show).   32. reservation\_status\_date: Date at which the last reservation status was updated.   **Data Cleaning:**  Data cleaning is done to ensure that the dataset is correct, consistent, and usable. It improves the efficiency and quality of analysis.  Data cleaning was done in 4 steps.   * Remove duplicate rows * Handling missing values. * Convert columns to appropriate datatypes * Adding impoertant columns.   **Exploratory Data Analysis:**  After cleaning of dataset , is subjected to exploratory data analysis, which will visualize the data and identify trends and patterns that can be later used to increase revenue.  EDA was carried out in 3 steps:   1. **Univariate Analysis** 2. **Bivariate Analysis** 3. **Correlation Analysis**   **Univariate Analysis**: A Uni means one and Variate means variable, so in univariate analysis, there is only one dependable variable. The objective of Uni-variate analysis is to derive the data, define and summarize it.  **Bivariate Analysis**: Bi means two and variate means variable, so here there are two variables. The analysis is related to cause and the relationship between the two variables.  **Correlation Analysis**: It is used to measure the strength of the linear relationship between two variables and compute their association. Correlation analysis calculates the level of change in one variable due to the change in the other.  **Challenges:**  Handling 119390 rows and 32 columns of data was a difficult as a beginner whithout any help. The inspection and cleaning of dataset was a time-consuming process. Visualization of data was properly carried out after providing a great amount of attention to each and every details. In one case outliers had to be removed to get a proper visualization and get the accurate result as needed.  **Conclusion:**  The following conclusions:   * City Hotel seems to be more preferred among travelers and it also generates more revenue. * Most number of bookings area made at cit hotel. * GDS channel brings higher revenue generating deals for City hotel. * TA/TO has highest booking cancellation %. |