Ticket price - Data Analysis & Predictive Modeling

Data Science Project

Project outcomes or summery

Step 1 Data Cleaning & Feature Engineering Summary:

- The data set shape changes from (10683, 11) to (8422 entries, 19 columns) after cleaning & Feature Engineering process.

- Maintain features columns name to be all lower case

- The summery of the cleaning & Feature Engineering for each feature as follows:

1. Date_of_Journey:

- Date_of_Journey data type changed from object to datetime.
- date_of_journey as date time it is not useful but we create new features from it like year, mon th, day, day name, then dropping the 'Date_of_Journey column.
- also, we create new features holiday with 1 refering to holiday and 0 for not holiday. after che cking from internet:

  * The standard working days in India is Monday to Friday then weekends days are Saturday an d Sunday

  * In India, National holidays vary according to its local state but there're ones which applied ov er the whole country like:

  - January 26 - > Republic Day Celebrates the 1950 adoption of the Constitution of India

  - August 15 - > Independence Day Celebrates the 1947 Independence from the British rule

  - October 2 - > Gandhi Jayanti Honors Mahatma Gandhi, father of the nation, who was born o n October 2, 1869

  - Reference: https://en.wikipedia.org/wiki/Public_holidays_in_India

- From the Month Features we Create New features Seasons. The climate of India consists of a wide range of weather conditions across a vast geographic scale and varied topography. but th e main seasons are:

  - Winter, occurring from December to February.

  - Summer or pre-monsoon season, lasting from March to May.

- Monsoon or rainy season, lasting from June to September.

- Post-monsoon or autumn season, lasting from October to November.

- Reference: https://en.wikipedia.org/wiki/Climate_of_India#Seasons

2. `destination`:

- We noticed that New Delhi and Delhi are both referring to the same destination with the same route abbreviation DEL. I replace New Delhi with Delhi value.

3. `dep_time`:

- split dep_time for two new columns one for hours and one for minutes then dropping it.
- Also, Creat dep_time as part of the day time like morning, afternoon, Evening, Night from dep_hour

4. `arrival_time`:

- We split arrival time to hour, minute, day, and month columns and then drop the arrival time.
- Also, we check if the arrival day not before departure day and deal with it by dropping 2037 indices of corrupted departure and arrival month and date.
- dropping arrival day and arrival month columns as they contain over 60% null values.
- furthermore, Create arrival time as part of the day time like (Early morning, morning, afternoon, Evening, Night, Late Night) from arrival hour.

5. `duration`:

- We found that there is strange value for duration of 5m we drop it.
- Also change duration value from hour and minutes to integer minutes values.

6. `additional_info`:

- There is typing issues like 'No info' and 'No Info' we handle it by changing all values to lower case.
- Almost 80% of the additional_info columns are no info or null values so we dropped it.

7. `total_stops`:

- when comparing stops obtaind from route column to total_stops:
  - minimum stops optain from route = 0 which is same as no-stop from total_stops.
  - maximum stops optain from route = 3 which is same as 3 stops from total_stops.
- from that we can asume that there is no contradictory between route & total_stops.
- Also, changing the values of total stops to numerical values.

- saving file as cleaned_data.csv after cleaning & Features Engineering process for next project step.

Step 2 Analysis outcomes or summary:

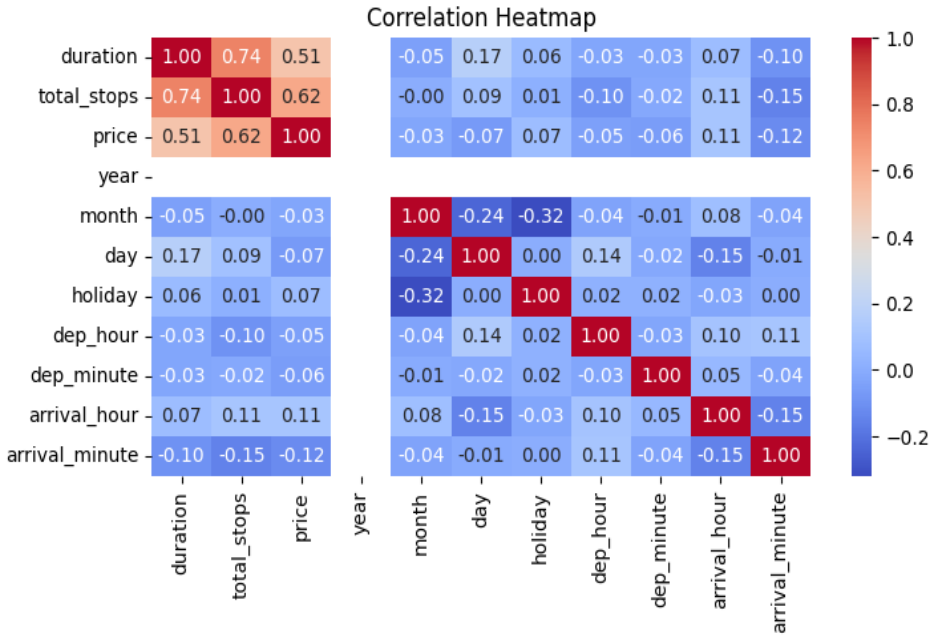
1. Statistical analysis:

- Explore important numerical and categorical features.

- **First Statistical Description for Some Numeric Features:**
    1. **price:** Ranges from **1759** to **79512**, The mean price is **8446.46**, and the median is **7318**. As the mean is greater than the median, it reflects that the data is right-skewed, and there are outliers in the upper ticket price values.
    2. **duration:** Ranges from **75** to **2820** minutes, The mean duration is **527.90**, and the median is **390.00**. As the mean is greater than the median, it reflects that the data is right-skewed, and there are outliers in the upper ticket price values.
    3. **year:** As all description like the min,max,median, and mean all have the same value 2019 which reflect all the data collected or concerned with year 2019.
- **Second Statistical Description for Categorical Features:**
    1. **airline:** Out of **8,422** airline ticket, the most frequented airline name is **Jet Airways**, which has **2575** values out of 12 total airline company unique name.
    2. **source:** Out of **8,422** airline ticket, the most frequented source city is **Delhi**, which has **3511** source values out of 5 total source city unique name.
    3. **destination:** Out of **8,422** airline ticket, the most frequented destination city is **Cochin**, which has **3511** values out of 5 total destination city unique name.
    4. **route:** Out of **8,422** airline ticket, the most frequented route is **DEL → BOM → COK**, which has **1924** values out of 112 total route unique name with only One stop trip.
    5. **day_name:** Out of **8,422** airline ticket, the most frequented day_name is **Thursday**, which has **1795** values out of 7 total unique day_name.
    6. **season:** Out of **8,422** airline ticket, the most frequented season is **Summer**, which has **4396** values out of 3 total unique seasons in the data indicate one missing season from the data and the data not covering all the year 2019.
    7. **dep_time:** Out of **8,422** airline ticket, the most frequented dep_time is **Morning**, which has **2365** values out of 6 total unique dep_time during the day.
    8. **arrival_time:** Out of **8,422** airline ticket, the most frequented arrival_time is **Evening**, which has **2296** values out of 6 total unique arrival_time during the day.

2. Correlation & Correlation Heatmap:

Correlation Analysis Summary



The correlation reveals the relationships between various features in the dataset:

1. Duration and Total Stops:

- Strong positive correlation (0.74): Longer durations are associated with more total stops.

2. Duration and Price:

- Moderate positive correlation (0.51): Longer durations are associated with higher prices.

3. Total Stops and Price:

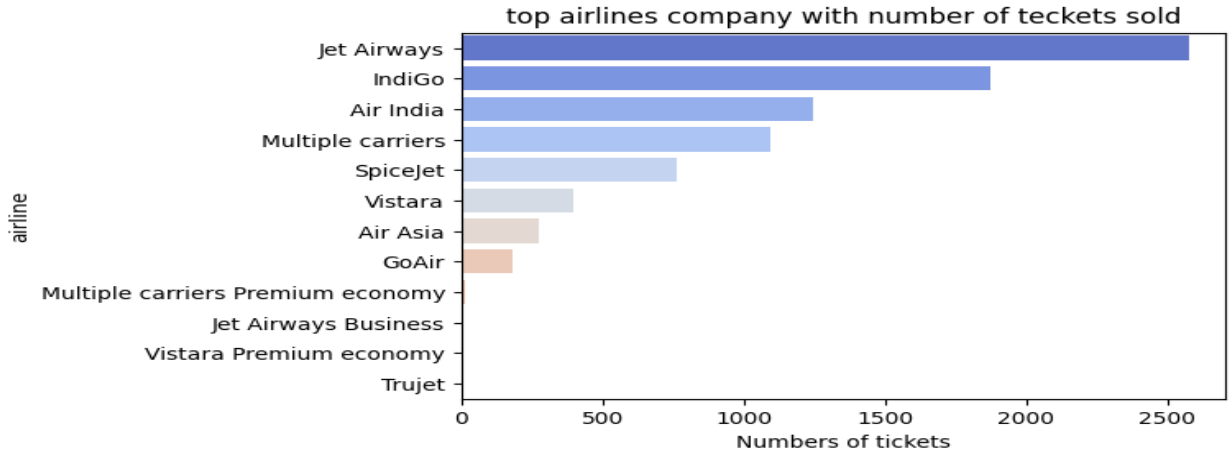
- Moderate positive correlation (0.62): More total stops are associated with higher prices. While it is often assumed that an increase in the number of total stops or longer flight durations would lead to lower ticket prices, the data suggests the opposite. The correlations indicate that, on average, longer durations and more total stops tend to be associated with higher ticket prices. This Unexpectedly finding underscores the complexity of factors influencing airfare and emphasizes the importance of considering various aspects when predicting ticket prices.

Note: Correlation does not imply causation, and these interpretations are based on the observed associations in the data.

3. Analysis and visualization around the following questions

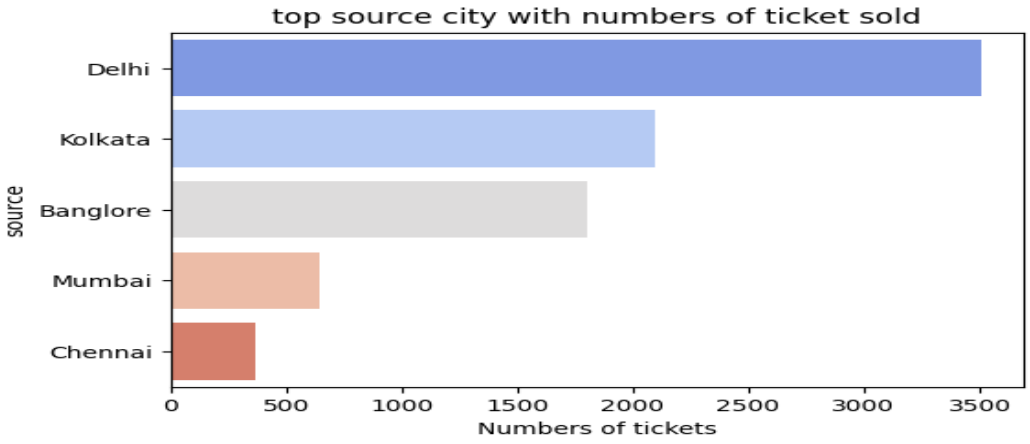
- Univariate analysis:

1. What are the top airline company with numbers of ticket sold?



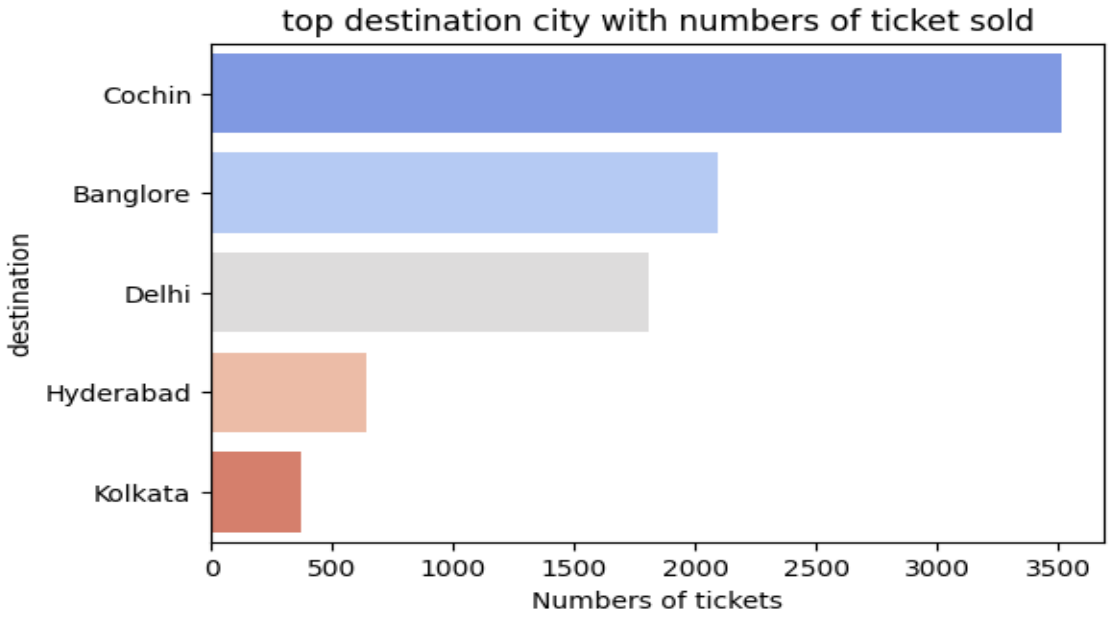
- The data shows Jet Airways as the most frequent airline (2575 instances), followed by IndiGo (1875) and other notable low-cost carriers. Legacy carrier Air India is also significant (1243 instances), with instances of multiple carriers (1093 occurrences). Premium economy options from Vistara and multiple carriers are mentioned, though less frequently (3 and 13 instances).

2. What are the top source city with numbers of ticket sold?



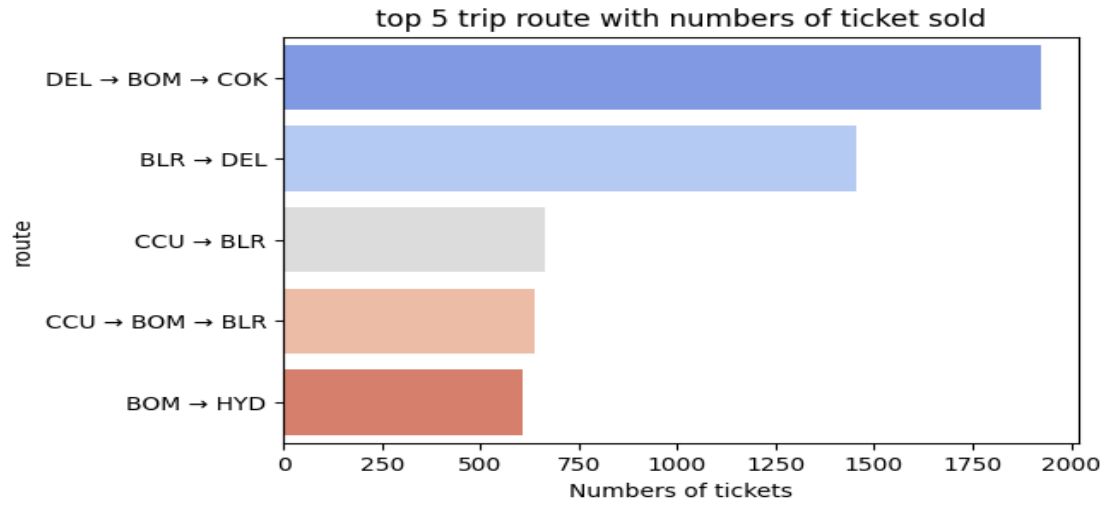
- The data displays the frequency distribution of source cities in the dataset, with Delhi being the most frequent (3511 instances), followed by Kolkata (2094) and Banglore (1806). Mumbai and Chennai also appear, though with lower frequencies (642 and 369 instances, respectively).

3. **What are the top destination city with numbers of ticket sold?**



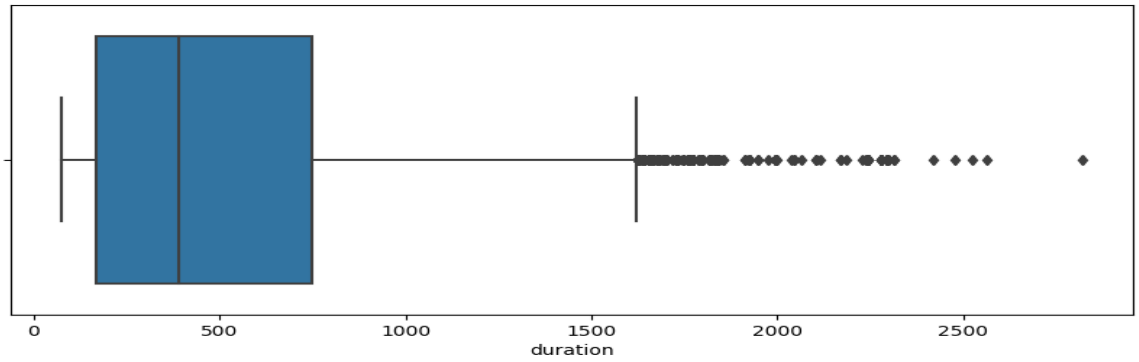
- The data illustrates the frequency distribution of destination cities in the dataset. Cochin is the most common destination (3511 instances), followed by Banglore (2094) and Delhi (1806). Hyderabad and Kolkata also appear, though with lower frequencies (642 and 369 instances, respectively).

4. **What are the top 5 route with numbers of ticket sold?**



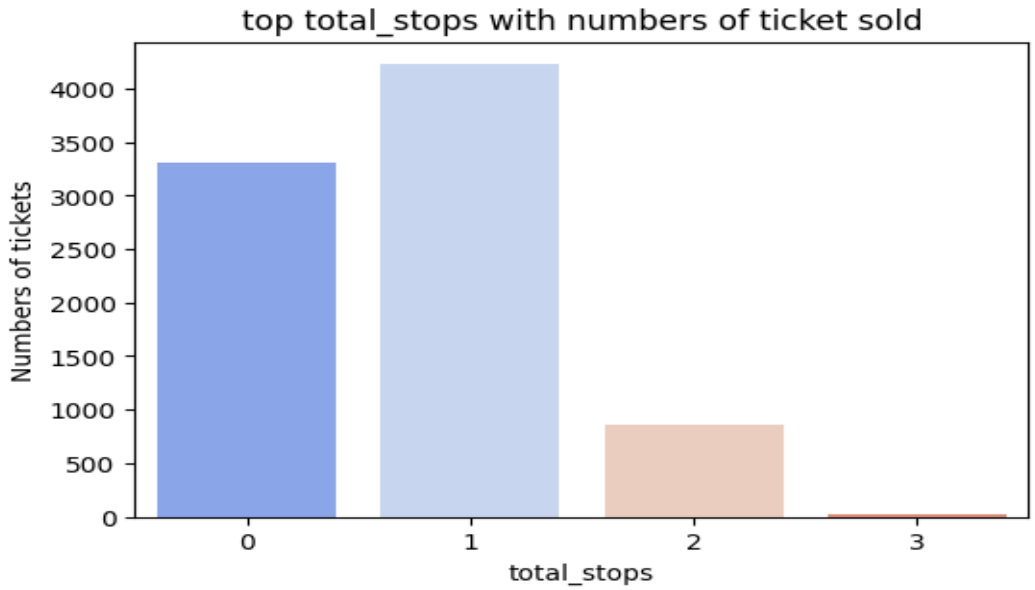
- The data showcases the frequency distribution of the top 5 trip routes in the dataset. The route DEL → BOM → COK is the most common, appearing 1924 times, followed by BLR → DEL (1455 instances) and CCU → BLR (665 instances). The route CCU → BOM → BLR and BOM → HYD also feature, though with slightly lower frequencies (637 and 607 instances, respectively).

5. **what is the mean duration and statistcal discription of the duration feature?**



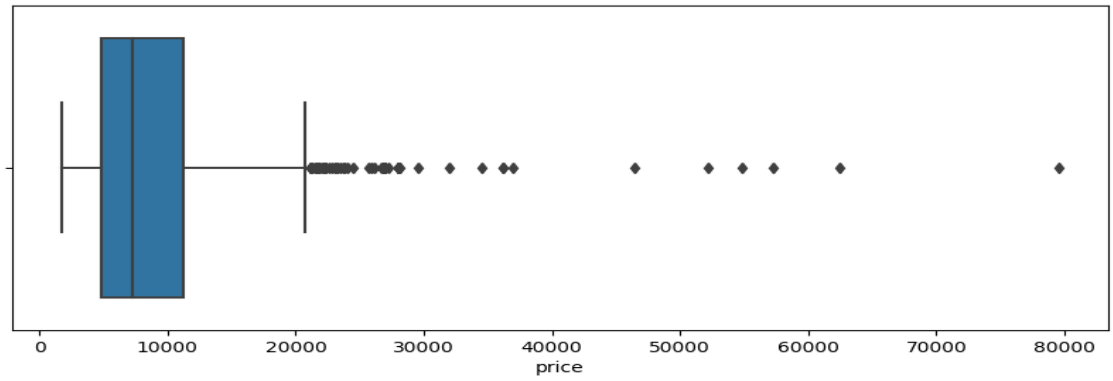
- The duration in minutes column statistics reveal key insights about the distribution of flight durations in the dataset. The data comprises 8422 instances, with a mean duration of approximately 527.90 minutes. The mean being greater than the median (50th percentile) suggests a right-skewed distribution, indicating that there are outliers contributing to higher durations. The variability is substantial, as indicated by the standard deviation of 447.17 minutes. The minimum duration is 75 minutes, with quartiles at 170 (25th percentile), 390 (50th percentile), and 750 (75th percentile) minutes. The maximum recorded duration is 2820 minutes.

6. **What is the top total_stops with numbers of ticket sold?**



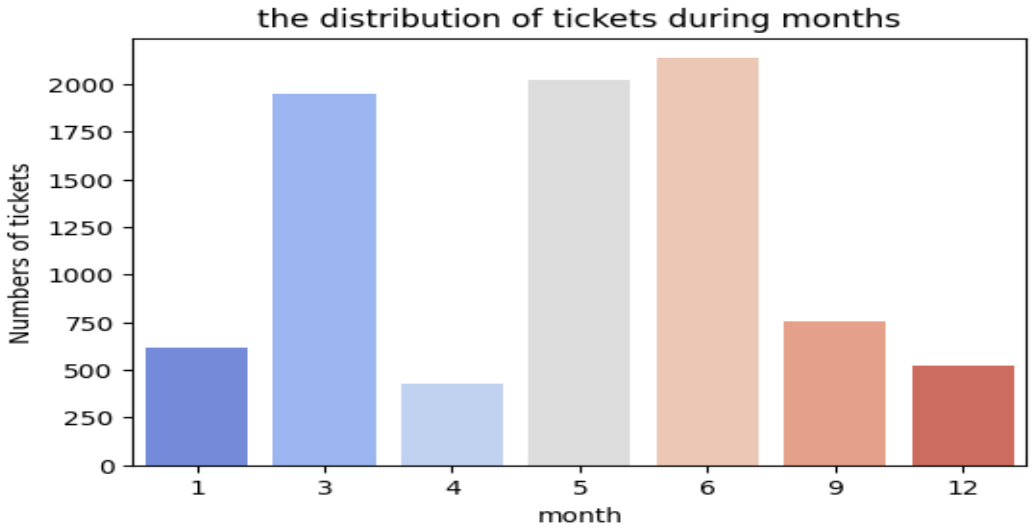
- The total trip stops distribution provides insights into the number of stops in the dataset. The majority of trips have 1 stop (4227 instances), followed by non-stop flights (3306 instances). Trips with 2 stops are less common but still significant (864 instances). Trips with 3 stops are relatively rare, with only 25 instances. This concise summary offers a quick overview of the prevalence of different total stop counts in the dataset, highlighting the dominance of single-stop and non-stop flights.

7. **what is the mean price and statistcal discription of the price feature?**



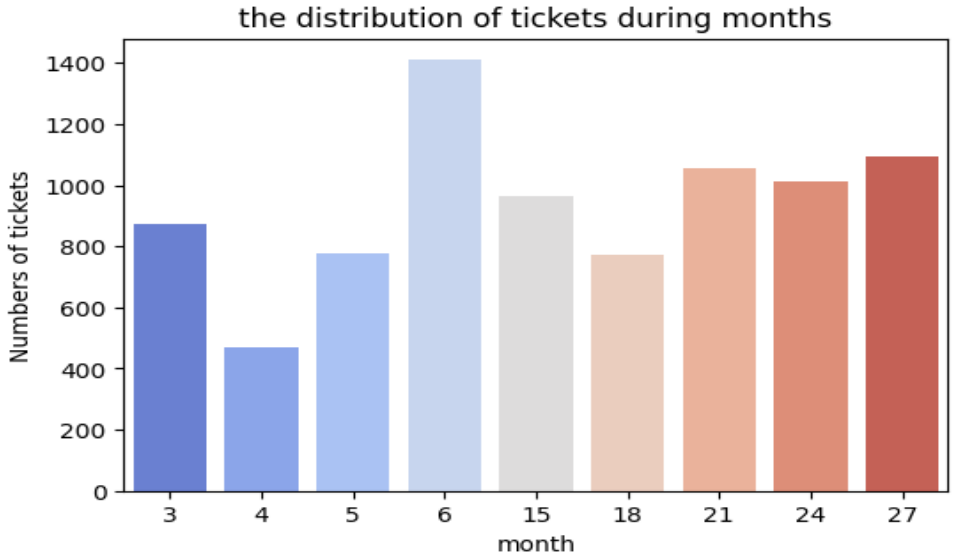
- The price column statistics reveal key insights about the distribution of ticket prices in the dataset. The data comprises 8422 instances, with a mean price of approximately ₹8446.46 (Indian Rupees). The fact that the mean is greater than the median (50th percentile) suggests a right-skewed distribution, indicating the presence of outliers contributing to higher ticket prices. The variability is substantial, as indicated by the standard deviation of ₹4415.85. The minimum ticket price is ₹1759, with quartiles at ₹4839.75 (25th percentile), ₹7318 (50th percentile), and ₹11264 (75th percentile). The maximum recorded price is ₹79512.

8. **What is the distribution of tickets during months?**



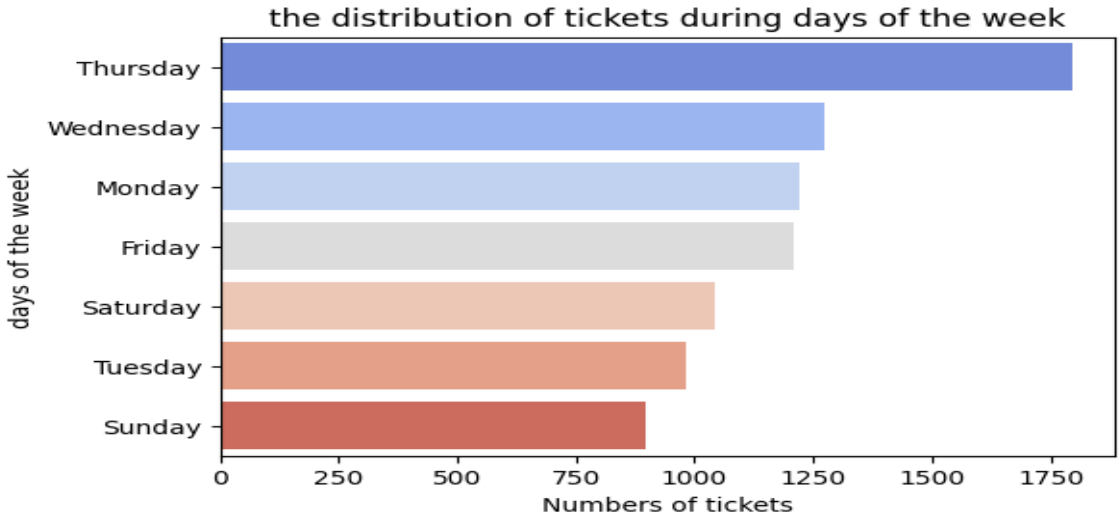
- The month table distribution highlights the frequency of flight occurrences in the dataset. June (Month 6) dominates with 2135 instances, followed by May (Month 5) at 2025, and March (Month 3) at 1947. September (Month 9) and January (Month 1) show lower frequencies with 752 and 619 instances, respectively. December (Month 12) and April (Month 4) have fewer instances, at 520 and 424.
- Notably, there is no data available for February, July, August, October, and November, indicating a lack of information for those months.

9. **What is the distribution of tickets during days of the month?**



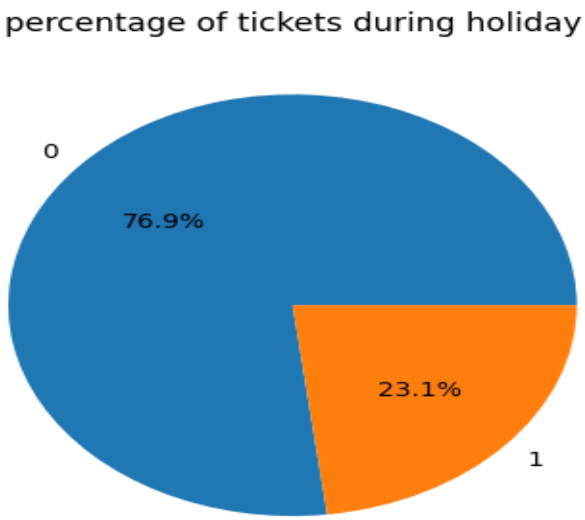
- The day of the month distribution reveals the frequency of flight occurrences on different days in the dataset. The 6th day has the highest frequency with 1409 instances, followed by the 27th (1092) and 21st (1054) days. The 24th and 15th days also show notable frequencies with 1013 and 962 instances, respectively. Days 3, 5, 18, and 4 have fewer instances, ranging from 468 to 778.

10. **What is the distribution of tickets during days of the week?**



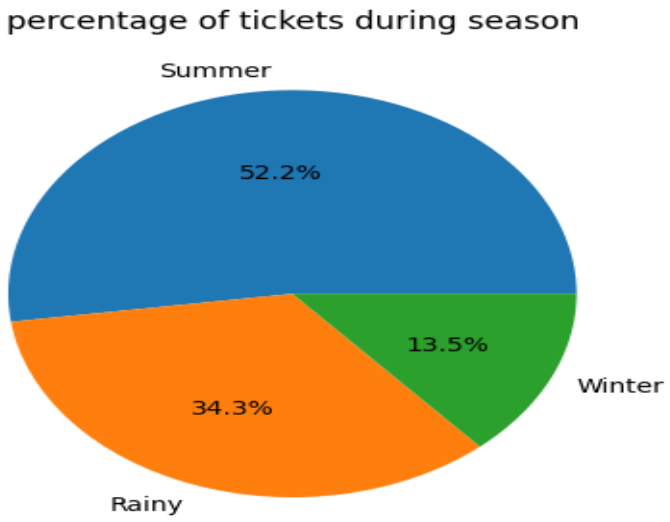
- The distribution of days of the week reveals the frequency of flight occurrences on different weekdays in the dataset. Thursday has the highest frequency with 1795 instances, followed by Wednesday (1274) and Monday (1222). Friday and Saturday also show notable frequencies with 1207 and 1044 instances, respectively. Tuesday and Sunday have fewer instances, with 982 and 898, respectively.

11. What is the distribution of tickets during holiday?



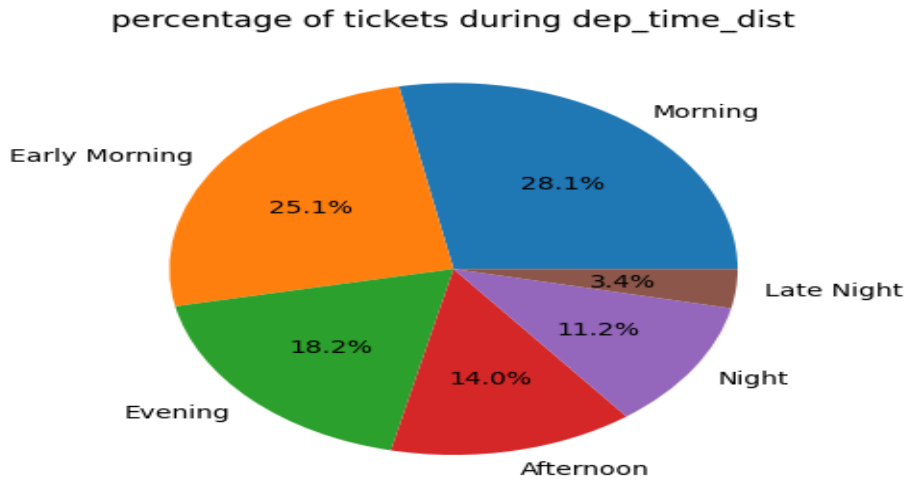
- The distribution of holiday values reveals the frequency of flight occurrences on holidays and non-holidays in the dataset. Non-holidays (value 0) have the highest 76.9% frequency with 6480 instances, while holidays (value 1) also show 23.1% a substantial occurrence with 1942 instances.

12. What is the distribution of tickets during seasons?



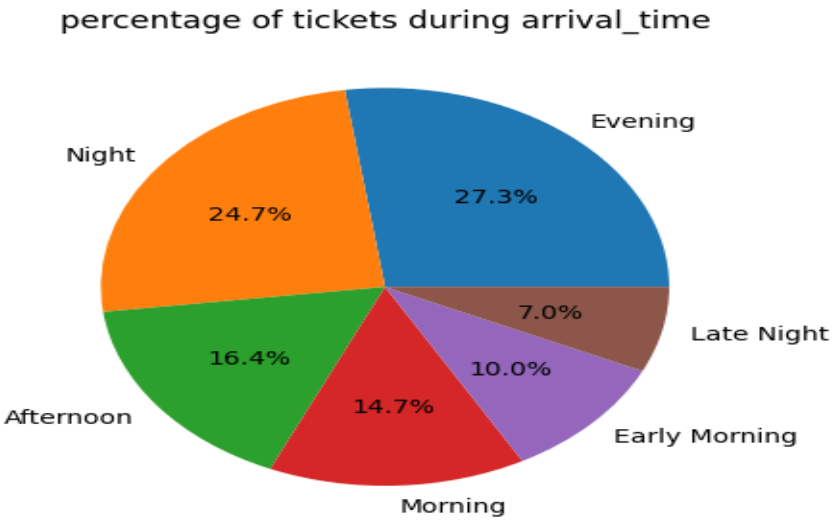
- The distribution of season values highlights the frequency of flight occurrences during different seasons in the dataset. 52.2% Summer has the highest frequency with 4396 instances, followed by 34.3% Rainy (2887) and 13.5% Winter with (1139) instances.
- As mentioned earlier the missing of October and November data reflect in missing Autumn season.

13. What is the distribution of tickets during dep_time?



- The distribution of departure times of the day provides insights into the frequency of flight occurrences during different time intervals. 28.1% Morning has the highest frequency with 2365 instances, followed by 25.1% Early Morning (2117) and 18.2% Evening (1536). 14% Afternoon and 11.2% Night also show notable frequencies with 1176 and 942 instances, respectively. Only 3.4% Late Night has a lower frequency with 286 instances.

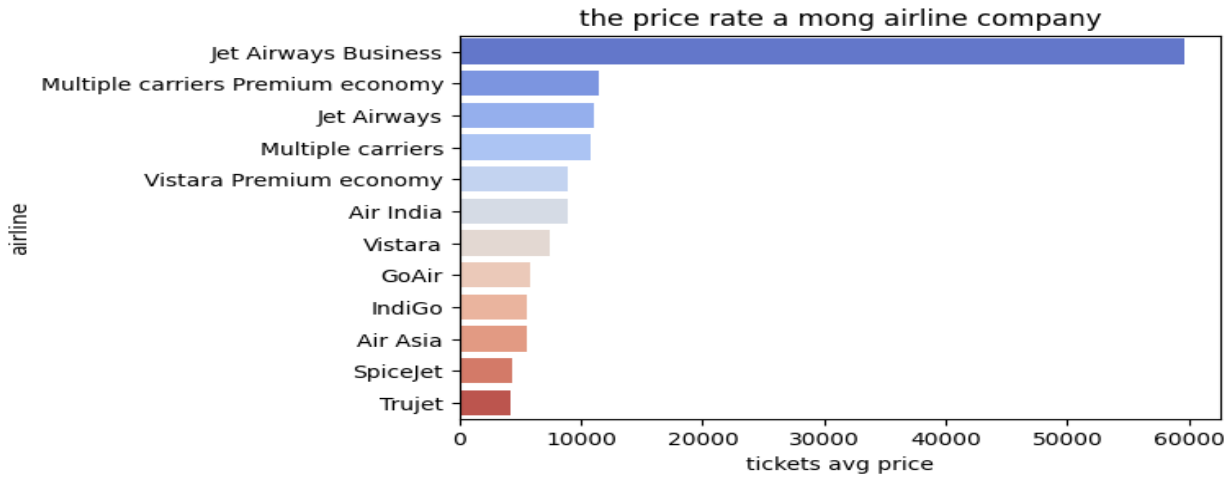
14. What is the distribution of tickets during arrival_time?



- The distribution of arrival times of the day sheds light on the frequency of flight occurrences during different time intervals. 27.3% Evening has the highest frequency with 2296 instances, followed by 24.7% Night (2084) and 16.4% Afternoon (1379). 14.7% Morning and 10% Early Morning also show notable frequencies with 1234 and 843 instances, respectively. 7% Late Night has a lower frequency with 586 instances.

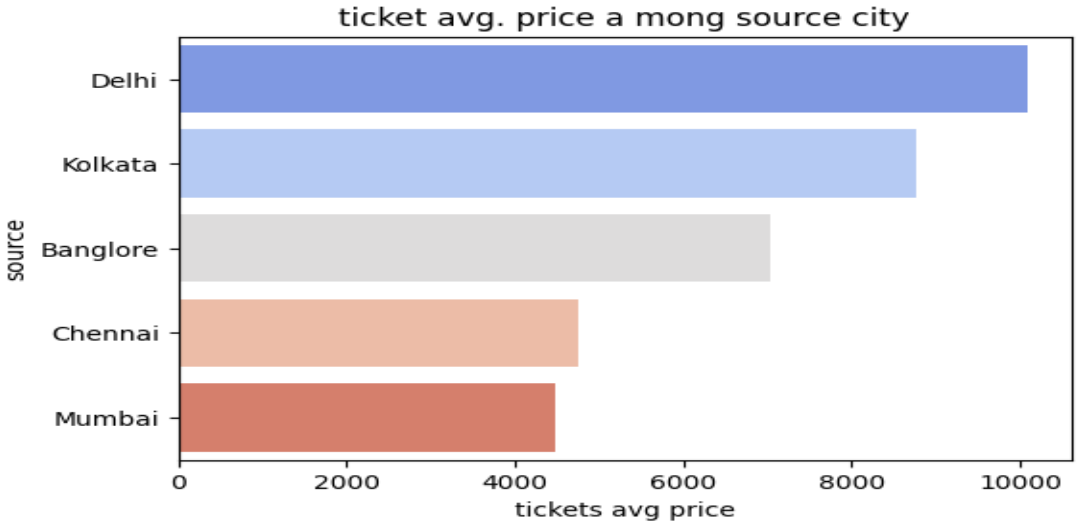
- Bivariate analysis:

1. Which airline company have higher mean price ticket?



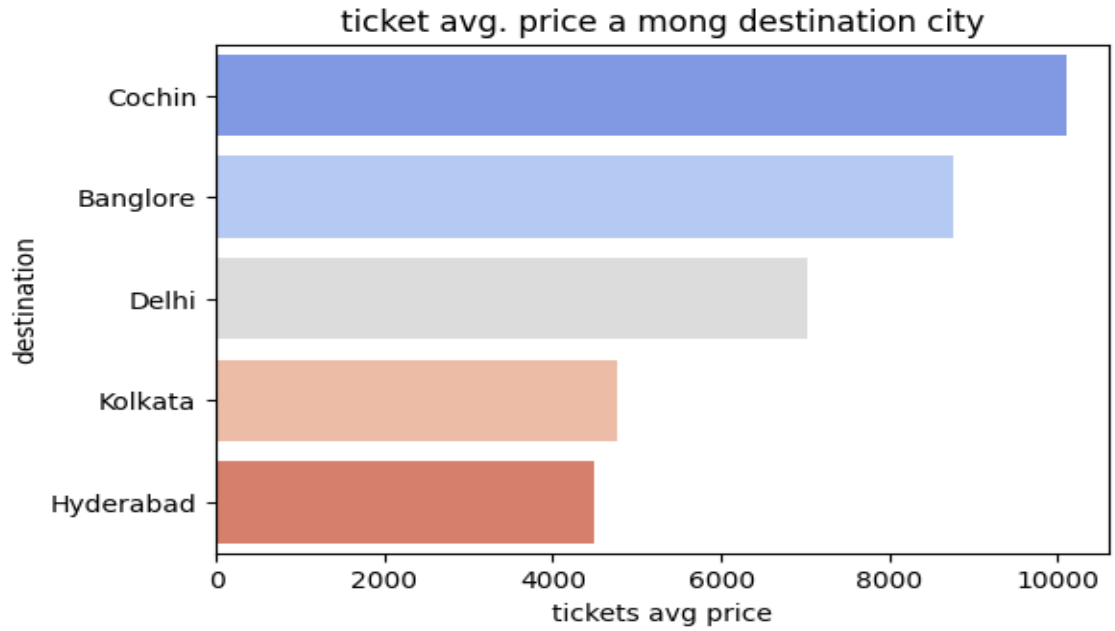
- The mean ticket prices for different airline companies provide insights into the average cost of flights with each carrier. Jet Airways Business has the highest mean price at ₹59573.40, followed by Multiple carriers' Premium economy (₹11418.85), Jet Airways (₹11058.11), and Multiple carriers (₹10753.92). Vistara Premium economy, Air India, and Vistara also have notable mean prices at ₹8962.33, ₹8916.58, and ₹7463.21, respectively. The mean prices gradually decrease for GoAir, IndiGo, Air Asia, SpiceJet, and Trujet.

2. Which source city have higher mean price ticket?



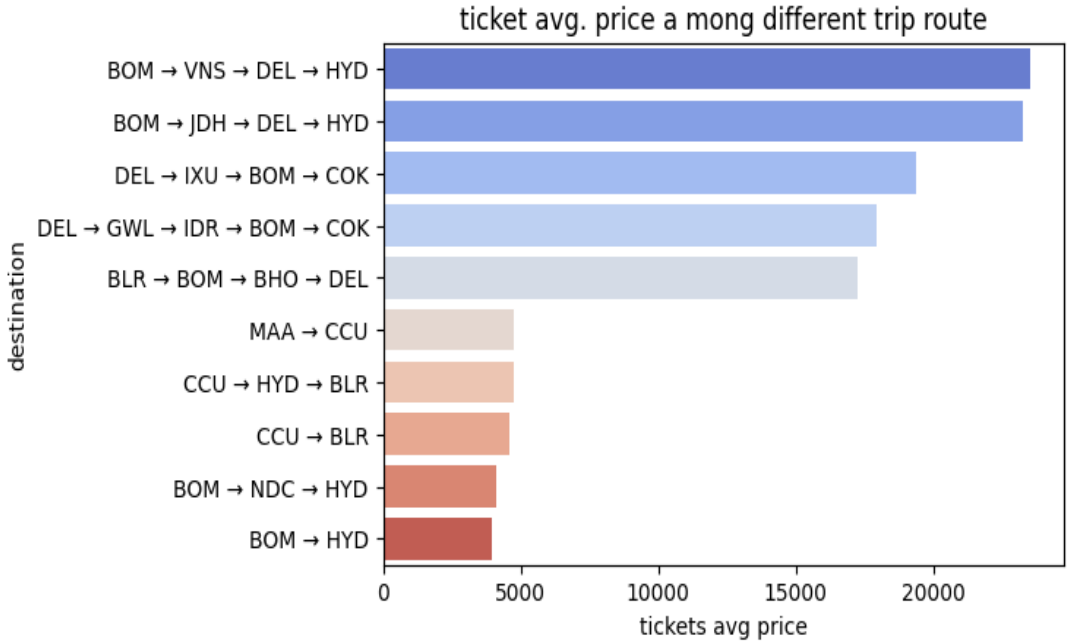
- The mean ticket prices for different source cities provide insights into the average cost of flights originating from each location. Delhi has the highest mean price at ₹10100.14, followed by Kolkata (₹8761.80), Banglore (₹7024.84), Chennai (₹4765.76), and Mumbai (₹4488.85).

3. **Which destination city have higher mean price ticket?**



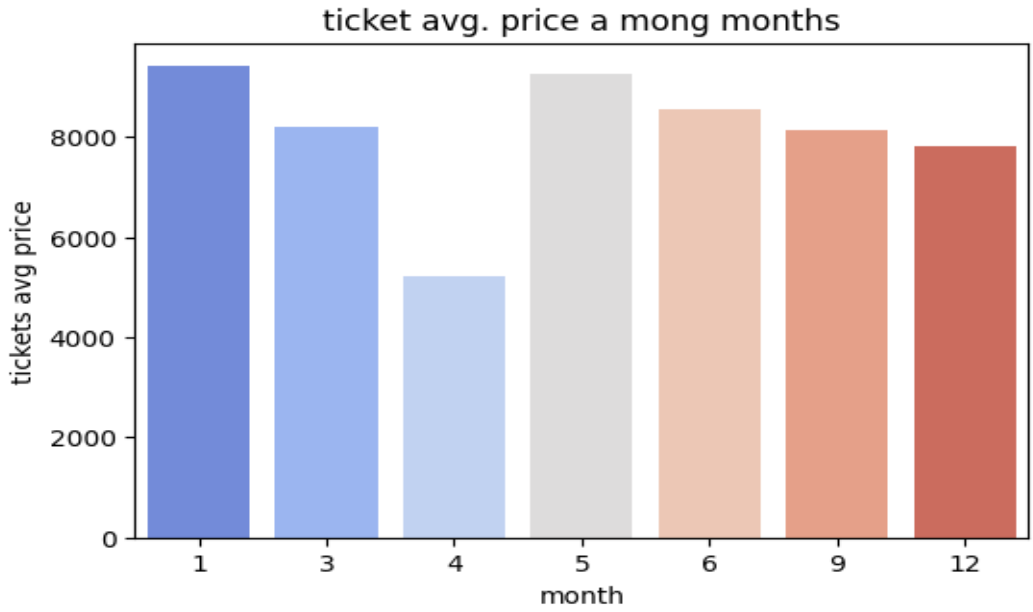
- The mean ticket prices for different destination cities provide insights into the average cost of flights arriving at each location. Cochin has the highest mean price at ₹10100.14, followed by Banglore (₹8761.80), Delhi (₹7024.84), Kolkata (₹4765.76), and Hyderabad (₹4488.85).

4. **Which trip route have higher mean price ticket?**



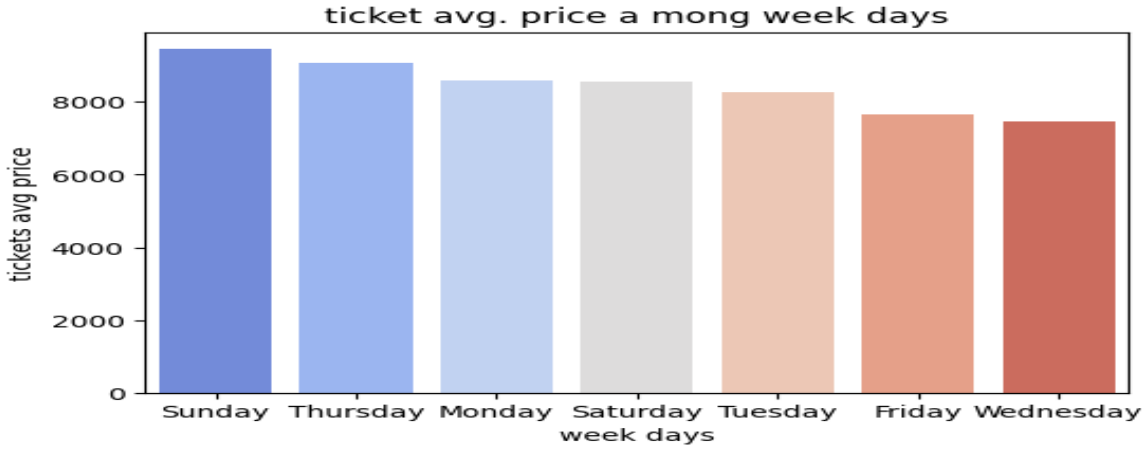
- the highest avg. price route with 2 stops BOM → VNS → DEL → HYD cost 23528 avg. price.
- the least avg.price route dirct or non-stop trip BOM → HYD with 3959.82 avg price cost.

5. **Which month have higher mean price ticket?**



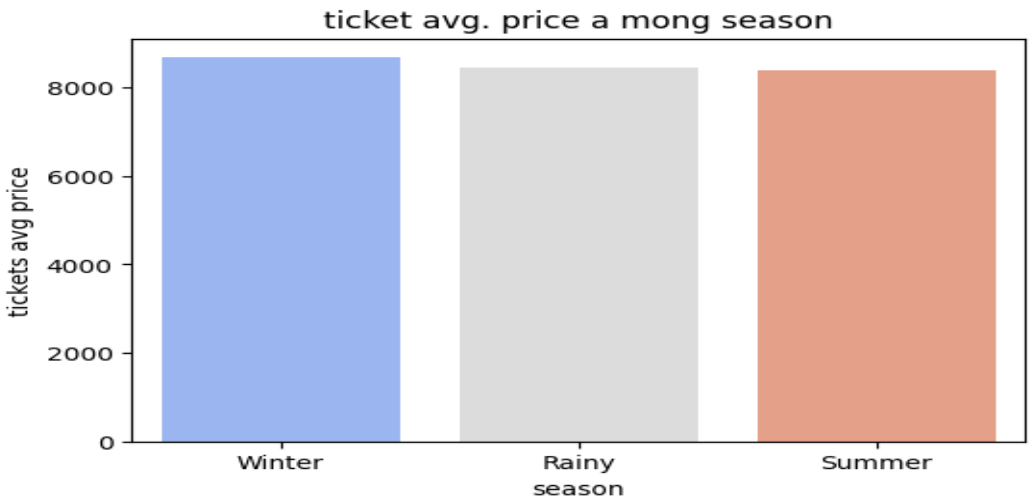
- The average ticket prices vary across different months, providing insights into seasonal trends. January sees a mean ticket price of ₹9406.10, while May follows closely at ₹9253.13. June maintains a slightly lower average at ₹8541.82, and March and September have average prices of ₹8186.92 and ₹8139.87, respectively. December shows a mean ticket price of ₹7808.33, and April has a comparatively lower average at ₹5230.86.

6. **Which week day have higher mean price ticket?**



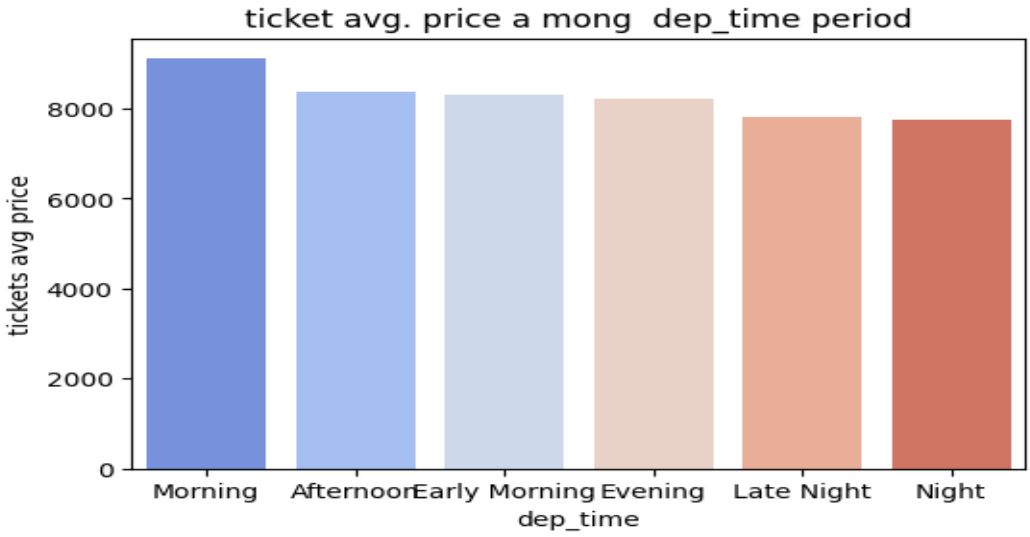
- The average ticket prices during different weekdays reveal fluctuations in pricing throughout the week. Sunday leads with a mean ticket price of ₹9448.87, followed by Thursday at ₹9080.84. Monday and Saturday maintain similar average prices, with ₹8596.31 and ₹8567.13, respectively. Tuesday shows a slightly lower average at ₹8276.37, while Friday has a mean ticket price of ₹7660.79. Wednesday marks the lowest average price among weekdays at ₹7478.91.

7. **Which season have higher mean price ticket?**



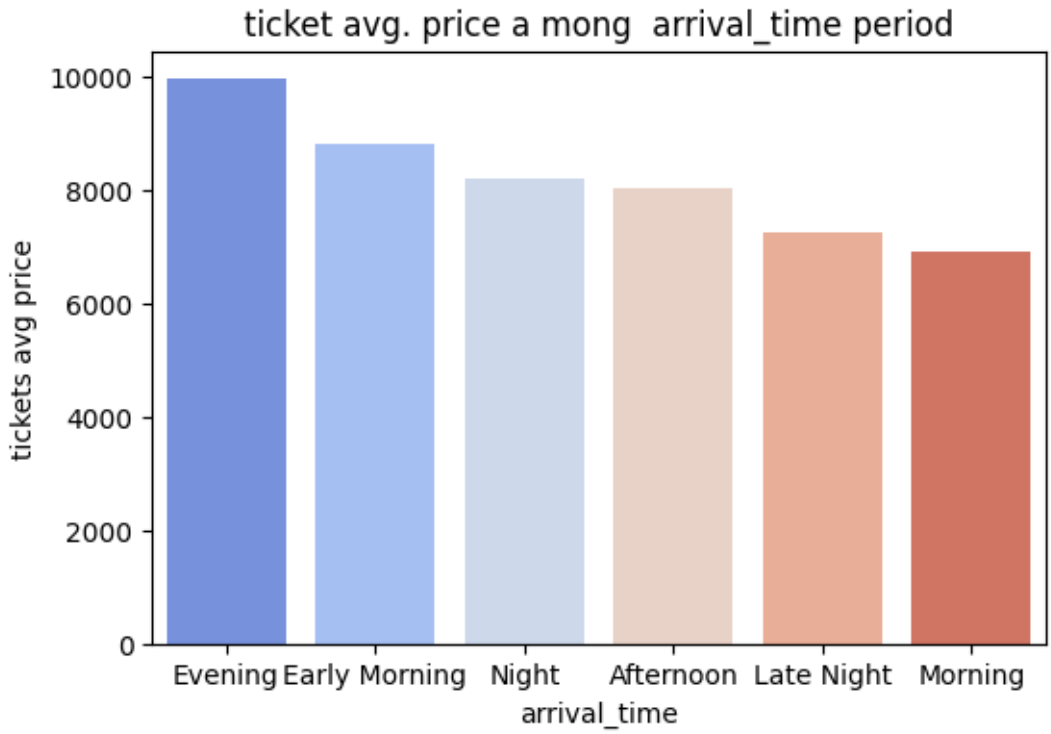
- The average ticket prices during different seasons highlight variations in pricing across seasonal periods. Winter exhibits the highest mean ticket price at ₹8676.65, followed closely by Rainy with an average of ₹8437.12. Summer shows a slightly lower mean price at ₹8392.95.

8. **Which dep_time have higher mean price ticket?**



- The average ticket prices during different departure times of the day reveal variations in pricing across various time intervals. Morning flights have the highest mean ticket price at ₹9112.48, followed by Afternoon (₹8375.73) and Early Morning (₹8298.63). Evening flights show a slightly lower mean price at ₹8230.87, while Late Night and Night flights have averages of ₹7815.69 and ₹7737.90, respectively.

9. **Which arrival_time have higher mean price ticket?**



- The average ticket prices during different arrival times of the day demonstrate variations in pricing across various time intervals. Evening arrivals have the highest mean ticket price at ₹9947.94, followed by Early Morning (₹8787.97) and Night (₹8183.96). Afternoon arrivals show a slightly lower mean price at ₹8018.36, while Late Night and Morning arrivals have averages of ₹7251.87 and ₹6908.47, respectively.

Analysis Conclusion:

In conclusion, this analysis provides a comprehensive understanding of the airline dataset, aiding in strategic decision-making for airlines and travel agencies. The insights into pricing, popular routes, and temporal trends can guide marketing, scheduling, and pricing strategies. Additionally, the identification of data gaps underscores the importance of data completeness for robust analyses. Future studies could delve deeper into specific factors influencing pricing and customer choices, contributing to a more nuanced understanding of the aviation industry.

Step 4: Preprocessing & Modeling

Preprocessing & Modeling summery:

1. Removing useless or Irrelevant Columns and modify some features:

    - Checking the dataset for irrelevant columns that do not serve our prediction and looking for duplications

2. Data Splitting:

    - Dividing the data set to features (x) and the target (y)

3. Handling Categorical Variables:

        - Binary encoding produced the best results during modeling.

4. Encoder Creation and Evaluation:

    - A column transformer was used to create an encoder with different encoding types.
    - Binary encoder was found to yield the best model accuracy scores.

5. Pipeline Creation and Model Evaluation:

    - A pipeline was constructed incorporating encoders, scalers, and various models for evaluation.
    - XGBRegressor model has the best 94% train, 83% test scores, 1129.75 Mean Absolute Errror, and 3281954.61 Mean Squared Error.
    - Overfitting was observed, but XGBRegressor still yielded the best test performance among tested models.

8. Hyperparameter Tuning:

    - RandomizedSearchCV was employed for hyperparameter tuning to enhance model performance and generalization.
    - Optimal parameters for XGBRegressor: {'XGR__subsample': 0.8, 'XGR__n_estimators': 600, 'XGR__max_depth': 9, 'XGR__learning_rate': 0.01, 'XGR__colsample_bytree': 0.8}
    - Best score achieved:

    - Best R-squared: 0.8408466505242934
    - Best MAE: 1083.5949883582057
    - Best MSE: 3089763.2108200593

- joblib was used to save the best model, along with column names.
- Streamlit online link for running app: https://indian-airline-ticket-prices-prediction-g8p5qcfqsemuzbfxxrcalz.streamlit.app/

Conclusion:

The data preprocessing and modeling steps aimed to create a robust model to predict Indian Airline Ticket price prediction. We focused on relevant features and utilized binary encoding for categorical variables, achieving optimal results. The XGBRegressor emerged as the best-performing model with an average test score of 83%. Despite indications of overfitting, it provided the most favorable test results. Hyperparameter tuning using RandomizedSearchCV improved the model's performance, resulting in a 84% best score. The model, along with its best parameters, was saved using joblib. The preprocessing steps, feature engineering, and model development contribute to a comprehensive framework for predicting Airline Ticket price in India.