Dataset Info:

1. Total records - 336776
2. Datatypes: 5 float, 9 int, 5 object type.

Null Values/ Missing Values:

1. Technical Analysis-
2. Tail number of **2512** flight records are unknown, tail number is given by FAA for every plane when they are sold and permitted to fly. So null value in tailnum shows that the **flight is discontinued**. Therefore, we can say that 2512 may be discontinued.
3. If tailnum is null then dep\_delay, arr\_delay, arr\_time, air\_time and dep\_time is null too. So, we can subtract these 2512 from every null field and focus on other null values for further missing value analysis.
4. Now, dep\_time null values are **5743** that indicates **cancelled flights**. And there are no rows with dep\_time and dep\_delay not being null together. So, it makes sense to say the cancelled flights didn't have dep\_delay as well. dep\_time and dep\_delay are highly related to each other.
5. When dep\_time is null, arr\_time is null too, that means we can assume that they are related and focus on flights with null arr\_time excluding flights having null values common in dep\_time and arr\_time fields
6. **458** flights have null arrival time and these flights were departed but not reached the destination may be because they were diverted or a technical error in database storage system. The perfect reason is unknown.
7. Thus, **717 flights** doesn't have their arrival delay reported due to unknown situation/circumstances. May be because of unpredictable conditions like Bad Weather, Security Check, External Threats, etc
8. Factual Analysis: (Just a possibility based on null values)
9. Discontinued Flights – **2512**
10. Cancelled Flights – **5743**
11. Diverted/Unknown Issue flights – **458**
12. Unreported Flights – **717**

Categories:

1. Factual Analysis:
2. Carrier with highest number of flights is UA (United Airlines) - **58665** flights
3. Carrier with lowest flights is OO (Skywest Airlines) with **32** flights
4. Total carriers in dataset – 16

Carrier Flights

* 1. UA 58665
  2. B6 54635
  3. EV 54173
  4. DL 48110
  5. AA 32729
  6. MQ 26397
  7. US 20536
  8. 9E 18460
  9. WN 12275
  10. VX 5162
  11. FL 3260
  12. AS 714
  13. F9 685
  14. YV 601
  15. HA 342
  16. OO 32

Distribution:

1. Destination:
2. Most Popular Destination according to the data - **Chicago O'Hare International Airport (ORD)**
3. Least Popular Destination according to the data - JAC, MYR, HDN, BZN, MTJ, EYW, PSP, SBN, LEX, CHO, ANC, LGA
4. Origin:
   * + 1. The dataset has three NYC airports as origin, in that most popular is **EWR: Newark Liberty International Airport**

c. Departure Time

According to KDE Plot, it seems around (7 to 9) hour the probability density is near to 0.8-0.85; and around (16 to 19) hour the probability density is near to (0.73-0.65) decreasing as the evening advances. So, the probability density of these two-time frames makes it popular as any new departure is more likely to be in this range.

Therefore, we can roughly say that **morning 7am to 9am and evening 4pm to 7pm** is most **popular departure time** and people prefer to travel at this hour.

1. **Arrival Time**

**Most preferred arrival time** is around **10am in morning and 6pm - 10pm in evening.**

**Correlations:**

1. Departure delay and Arrival delay. It perfectly makes sense if a flight is departed at a delayed time, then it will arrive late too. But no direct/strong conclusion be made just based on heatmap.
2. Airtime and Distance are also highly related, as its obvious that more distance will take more airtime for the flight to travel.
3. Other fields like hour, departure time are also highly related but in actual it doesn't make much sense to focus on those.

**Studying Individual Relationship – Scatter Plots**

Here focusing on just 2 important relations we derived from heatmap : air\_time Vs. distance & arr\_delay Vs. dep\_delay

1. distance Vs. air\_time :

The scatter plot from above pairplot for this relation shows **directly proportional linear relation** which signifies that increase in distance increases air\_time too. This result is completely believable as more distance takes more time to travel.

1. dep\_delay Vs. arr\_delay :

The scatter plot for this relation also shows **directly proportional linear relation** that explains how delay in departure subsequently leads to delay in arrival time of a flight. It's agreeable to directly relate these two fields to each other.

**Conclusion: We can consider any one of the column from the relation and avoid another to reduce the duplicate calculation and interpretation of analysis(it reduces processing time and CPU power). For instance, we can just take distance and perform our analysis without considering air\_time and at last conclude our result for both fields because they are related.**

Outliers and Flight Frequency:

Four important information can be found here:

1. Most flights travel in between 0 to 2000 miles with few outliers travelling more than that around 3000miles and beyond 4000miles

2. Air time of majority of flights is within 2 hours with few flights having 4,5,6 hours of flight air time. The long hours flight indicates the flights with larger distance travelled.

3. Majority of flights occur between 10th to 25th date of every month.

4. Most of the flights were recorded in between April and October months of the year.

Missing Values Handling:

I have not deleted the rows with null values because from my point of view every row in this database signifies something.

The rows having null values in tailnum, arr\_time, arr\_delay, dep\_time, dep\_delay, air\_time signifies that these flights are discontinued and won't be used in future. It can give out crucial information of discontinued flights when we analyse business of flight carriers or check maintenance cost of airports with the help of other datasets.

According to me, the steps of deletion of rows or filling null values with interpolation, filling mean/median values, checking variance/Standard deviation doesn't make much sense in this dataset.

**Profile Report: An HTML file output.html can be viewed to see profile report.**