



FINAL PROJECT DATA SCIENCE & ANALYST  
STUDI INDEPENDEN BATCH 6

# FLIGHT PRICE PREDICTION



Mentor : Rusnanda Farhan





# KELOMPOK 5



M. Ilham Ramadhan  
2110511078



Fairuz Dwi Najla  
24050121130074



M. Andryan Maulana  
202131114



Putri Nabilla  
2106038



Misael Algospel  
222406034

# PENDAHULUAN



Kemajuan dalam industri pariwisata telah meningkatkan minat wisatawan untuk menjelajahi berbagai destinasi, memberikan dampak positif pada sektor penerbangan dan pariwisata. Teknologi telah mempengaruhi penjualan tiket pesawat, mendukung kemajuan industri ini. Data menjadi penting dalam menganalisis perilaku konsumen dan dinamika pasar, membantu perusahaan penerbangan memahami pola perilaku konsumen, menyesuaikan strategi harga, dan meningkatkan pengalaman pelanggan.

Deregulasi sektor penerbangan telah mendorong pengembangan kerangka aturan dan model penetapan harga tiket yang canggih. Maskapai harus mengelola ketersediaan tiket dengan cermat untuk mengoptimalkan pendapatan berdasarkan permintaan konsumen yang fluktuatif. Prediksi harga tiket pesawat menjadi sangat penting dalam mengatur harga secara optimal.

Penelitian ini menganalisis dataset pemesanan penerbangan dari situs "Ease My Trip" menggunakan metode Machine Learning seperti Regresi Linier, Random Forest, dan Neural Network. Dataset mencakup lebih dari 300.000 transaksi pemesanan penerbangan antara enam kota metro utama di India dengan 11 fitur seperti maskapai, kota asal dan tujuan, waktu keberangkatan, kelas tempat duduk, durasi perjalanan, dan harga tiket. Tujuannya adalah mendapatkan wawasan mendalam mengenai faktor-faktor yang mempengaruhi harga tiket.

# METODE PENELITIAN

Penelitian ini, dilakukan menggunakan media bantu Google Colaboratory untuk membuat sebuah model data dengan algoritma Regresi Linier, Random forest dan Neural Network. Penelitian diawali dengan pengambilan dataset dari website Kaggle yaitu dataset dari Flight Price Prediction milik Shubham Bathwal yang diperoleh dari situs Ease My Trip. Dataset ini terdiri dari 300.261 opsi pemesanan penerbangan berbeda diambil dari situs ini. Pengambilan data dilakukan selama 50 hari, mulai tanggal 11 Februari hingga 31 Maret 2022. Kumpulan data berisi informasi tentang opsi pemesanan penerbangan dari situs web Easemytrip untuk perjalanan penerbangan antara 6 kota metro teratas di India. Terdapat 30.0261 titik data dan 11 fitur dalam kumpulan data yang dibersihkan.

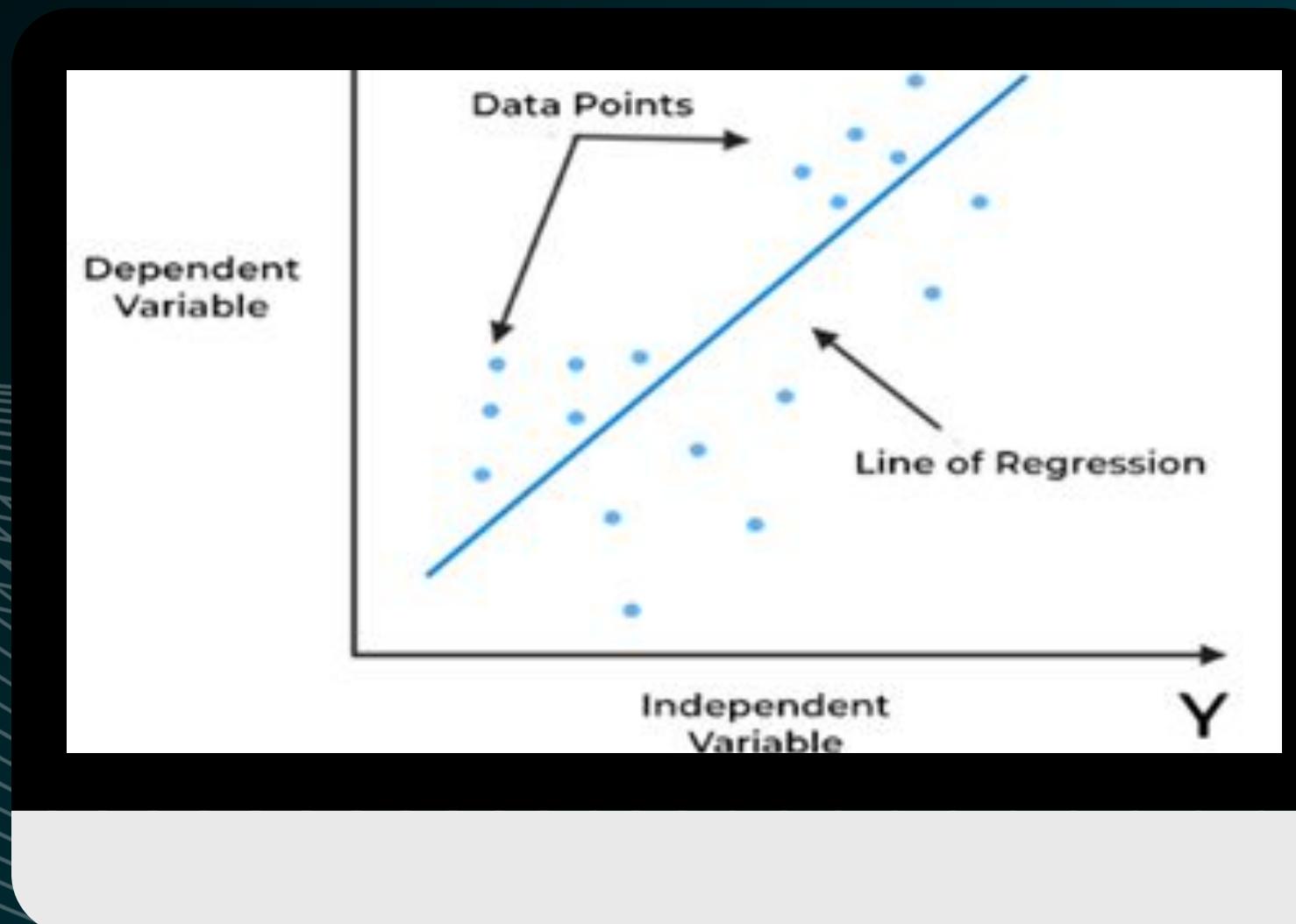
	Unnamed: 0	airline	flight	source_city	departure_time	stops	arrival_time	destination_city	class	duration	days_left	price
0	0	SpiceJet	SG-8709	Delhi	Evening	zero	Night	Mumbai	Economy	2.17	1	5953
1	1	SpiceJet	SG-8157	Delhi	Early_Morning	zero	Morning	Mumbai	Economy	2.33	1	5953
2	2	AirAsia	I5-764	Delhi	Early_Morning	zero	Early_Morning	Mumbai	Economy	2.17	1	5956
3	3	Vistara	UK-995	Delhi	Morning	zero	Afternoon	Mumbai	Economy	2.25	1	5955
4	4	Vistara	UK-963	Delhi	Morning	zero	Morning	Mumbai	Economy	2.33	1	5955

Setelah dataset diunduh dilakukan pada tahap Exploratory Data Analysis data yang di dalamnya terdapat Descriptive Analysis dan Data Visualization. Setelah itu dilakukan tahap Pre-processing yang di dalamnya terdapat Normalisasi Numerical untuk memastikan variable numerik memiliki skala yang seragam, Outliers untuk membersihkan atau menghapus outlier dari kolom – kolom numerik dalam sebuah DataFrame, Label Encoding untuk mengubah data kategori menjadi bentuk numerik, Code Correlation Matrix untuk mengevaluasi hubungan statistic antara pasangan variable numerik. Setelah tahapan Pre-Processing selesai, dilakukan tahapan Pembagian Data. Setelah itu dilakukan Modelling tahapan ini terdiri dari Predicting, Evaluating Model, Difference Between Actual & Predicted Price, Feature Importance. Tahapan terakhir yang dilakukan adalah Model Optimization tahapan ini terdiri dari Randomized search dan Grid Search.

# LINEAR REGRESSION



## Model Regresi Linear



## Pengertian

Menurut Gujarati (2006), analisis regresi linier adalah suatu analisis yang mempelajari hubungan ketergantungan antara satu variabel yang disebut variabel terikat terhadap variabel lain yang disebut variabel bebas. Dengan analisis regresi dapat diperhitungkan besarnya pengaruh dari perubahan satu variabel terhadap lain. Regresi linier pun dapat membentuk hubungan antara variabel bebas terhadap variabel terikat secara linier.



## Manfaat

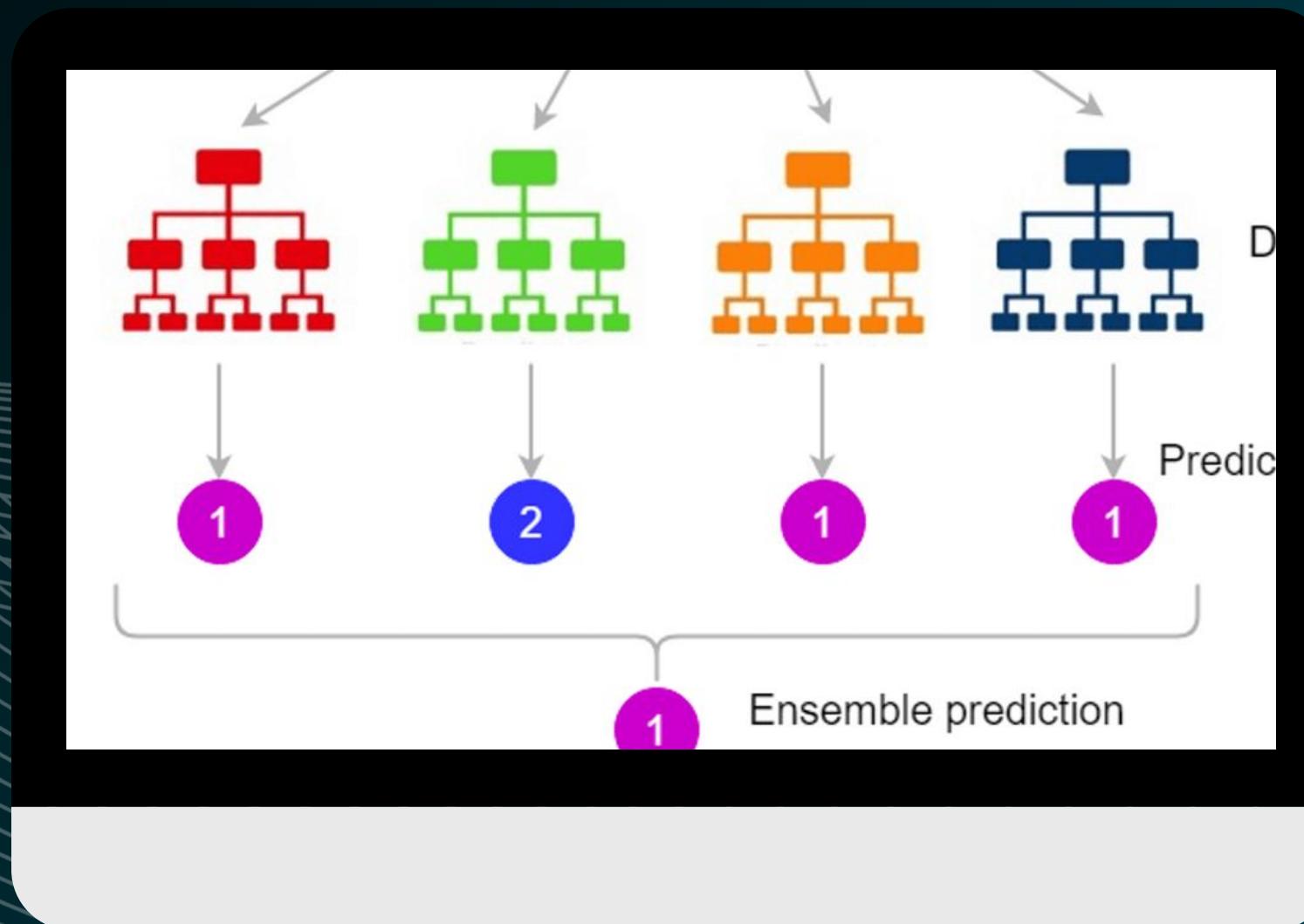
Algoritma Linear Regression memiliki beberapa manfaat yang signifikan dalam analisis data dan pembelajaran mesin. Berikut adalah beberapa di antaranya:

1. Implementasi yang mudah
2. Interpretasi
3. Skalabilitas
4. Optimal untuk pengaturan online

# RANDOM FOREST



## Random Forest



## Pengertian

Menurut Leo Breiman (2001) Random Forest adalah bentuk dari ensemble learning, di mana beberapa model (dalam konteks ini, pohon keputusan) digabungkan bersama untuk meningkatkan kinerja dan stabilitas prediksi. Popularitasnya berasal dari kemudahan penggunaan dan keserbagunaannya, sehingga cocok untuk tugas klasifikasi dan regresi.



## Manfaat

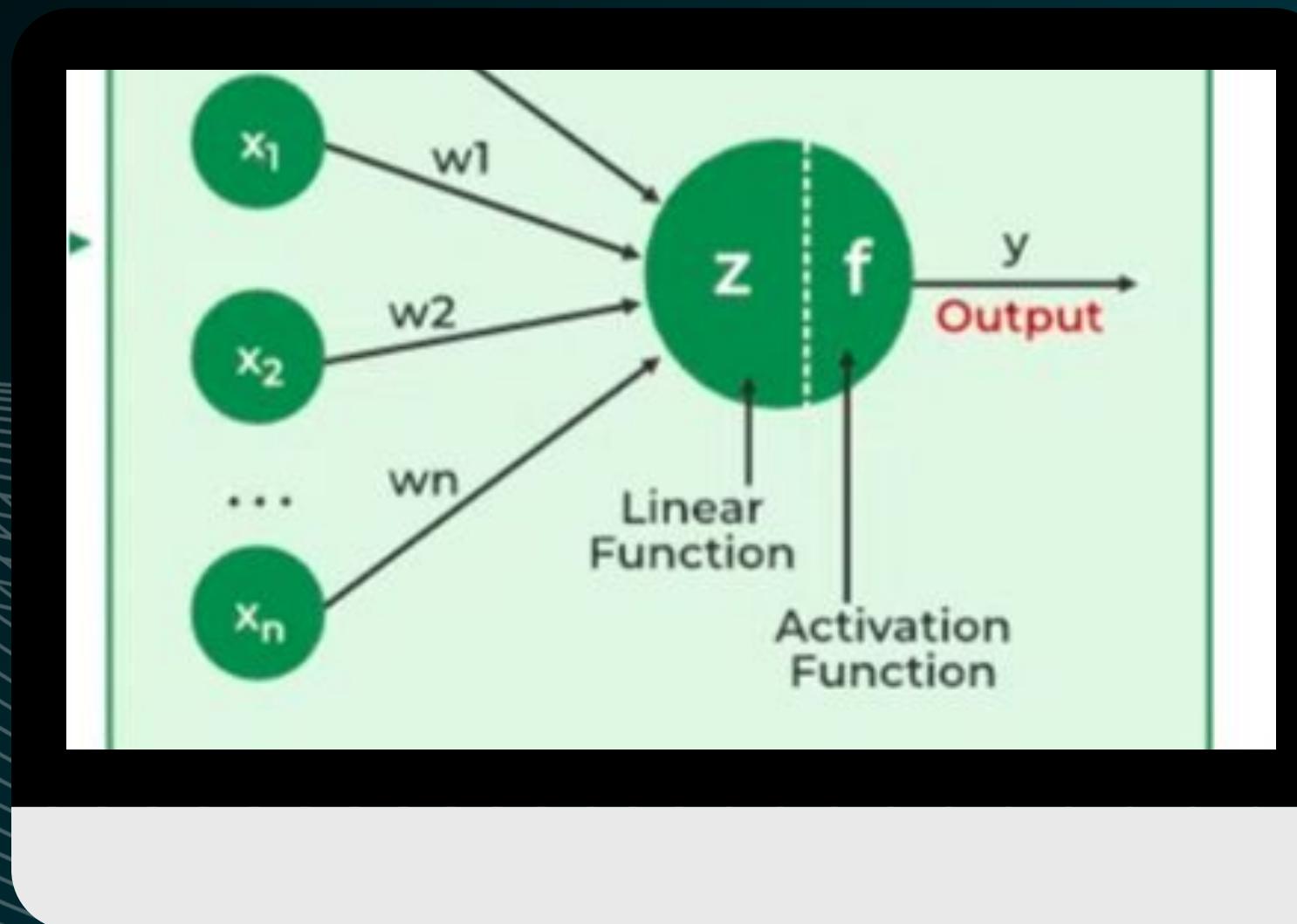
Algoritma Random Forest memiliki beberapa manfaat yang signifikan dalam analisis data dan pembelajaran mesin. Berikut adalah beberapa di antaranya:

1. Akurasi Tinggi
2. Mampu mengatasi Masalah Overfitting
3. Fleksibilitas dalam Tipe Data

# NEURAL NETWORK



## Neural Network



## Pengertian

Menurut Hinton (1986) Neural Network adalah sistem komputasi yang terinspirasi oleh jaringan saraf biologis di otak manusia, terdiri dari lapisan-lapisan neuron tiruan yang bekerja bersama untuk memproses informasi dan belajar dari data. Kemampuan neural network untuk mengidentifikasi pola, memecahkan teka-teki rumit, dan menyesuaikan diri dengan perubahan lingkungan kapasitas mereka untuk belajar dari data memiliki dampak yang luas



## Manfaat

Algoritma Neural Network memiliki beberapa manfaat yang signifikan dalam analisis data dan pembelajaran mesin. Berikut adalah beberapa di antaranya:

1. Kemampuan untuk Belajar dari Data
2. Skalabilitas
3. Kemampuan untuk Memodelkan Data Kompleks



# HASIL DAN PEMBAHASAN

	airline	flight	source_city	departure_time	stops	arrival_time	destination_city	class	duration	days_left	price
0	SpiceJet	SG-8709	Delhi	Evening	zero	Night	Mumbai	Economy	2.17	1	5953
1	SpiceJet	SG-8157	Delhi	Early_Morning	zero	Morning	Mumbai	Economy	2.33	1	5953
2	AirAsia	I5-764	Delhi	Early_Morning	zero	Early_Morning	Mumbai	Economy	2.17	1	5956
3	Vistara	UK-995	Delhi	Morning	zero	Afternoon	Mumbai	Economy	2.25	1	5955
4	Vistara	UK-963	Delhi	Morning	zero	Morning	Mumbai	Economy	2.33	1	5955

Gambar 1. Dataset

Dataset ini diambil dari situs web "Ease My Trip" yang merupakan platform untuk pemesanan tiket pesawat. Dataset ini mencakup informasi tentang opsi pemesanan penerbangan antara 6 kota besar di India, dengan total 300.261 baris data dan 11 kolom data. 11 kolom atau fitur untuk pemodelan ini meliputi : airline (nama maskapai), flight (kode penerbangan), source\_city (kota asal penerbangan), departure\_time (waktu keberangkatan), stops (jumlah pemberhentian atau transit), arrival\_time (waktu kedatangan), destination\_city (kota tujuan), duration (durasi perjalanan dalam jam), days\_left (sisa hari hingga keberangkatan) dan price (harga tiket).

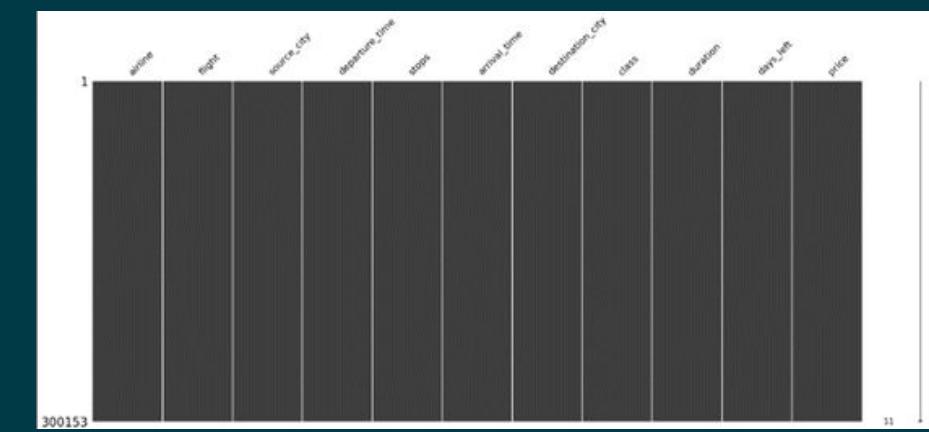
	Column Name	Number of Unique Values	Unique Values	Data Type
0	airline	6	-	object
1	flight	1561	-	object
2	source_city	6	-	object
3	departure_time	6	-	object
4	stops	3	[zero, one, two_or_more]	object
5	arrival_time	6	-	object
6	destination_city	6	-	object
7	class	2	[Economy, Business]	object
8	duration	476	-	float
9	days_left	49	-	int
10	price	12157	-	int

## Gambar 2. Analisis Deskriptif

```
df.isnull().sum()

airline
flight
source_city
departure_time
stops
arrival_time
destination_city
class
duration
days_left
price
dtype: int64
```

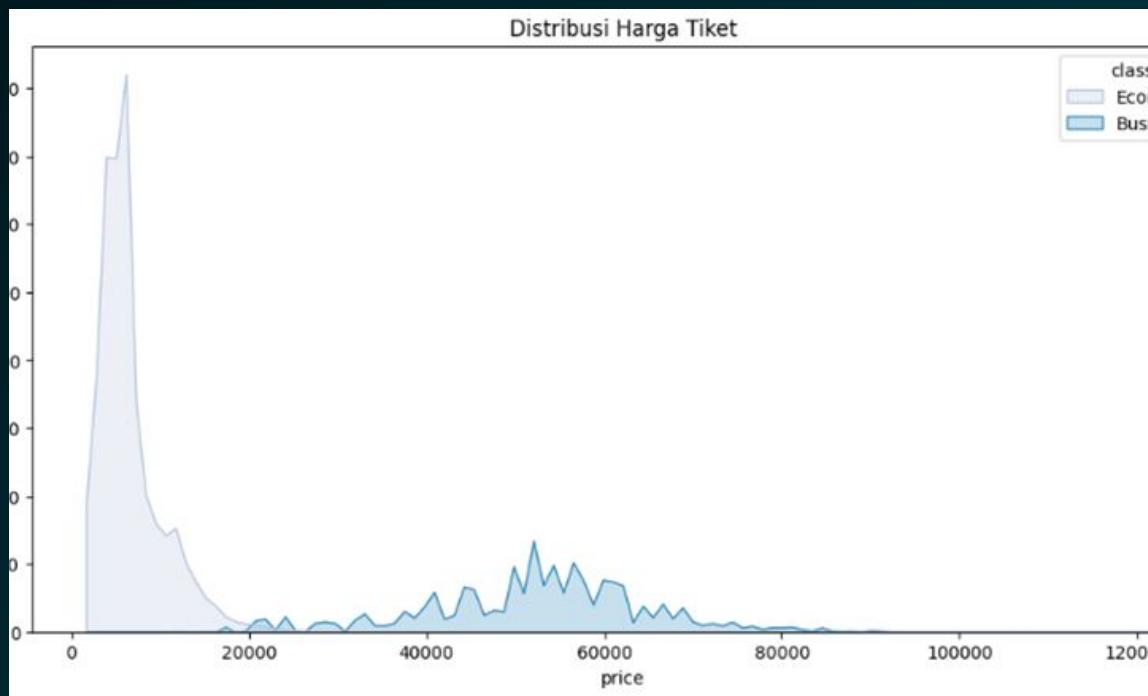
Gambar 3. Melihat Nilai Kosong Pada Dataset



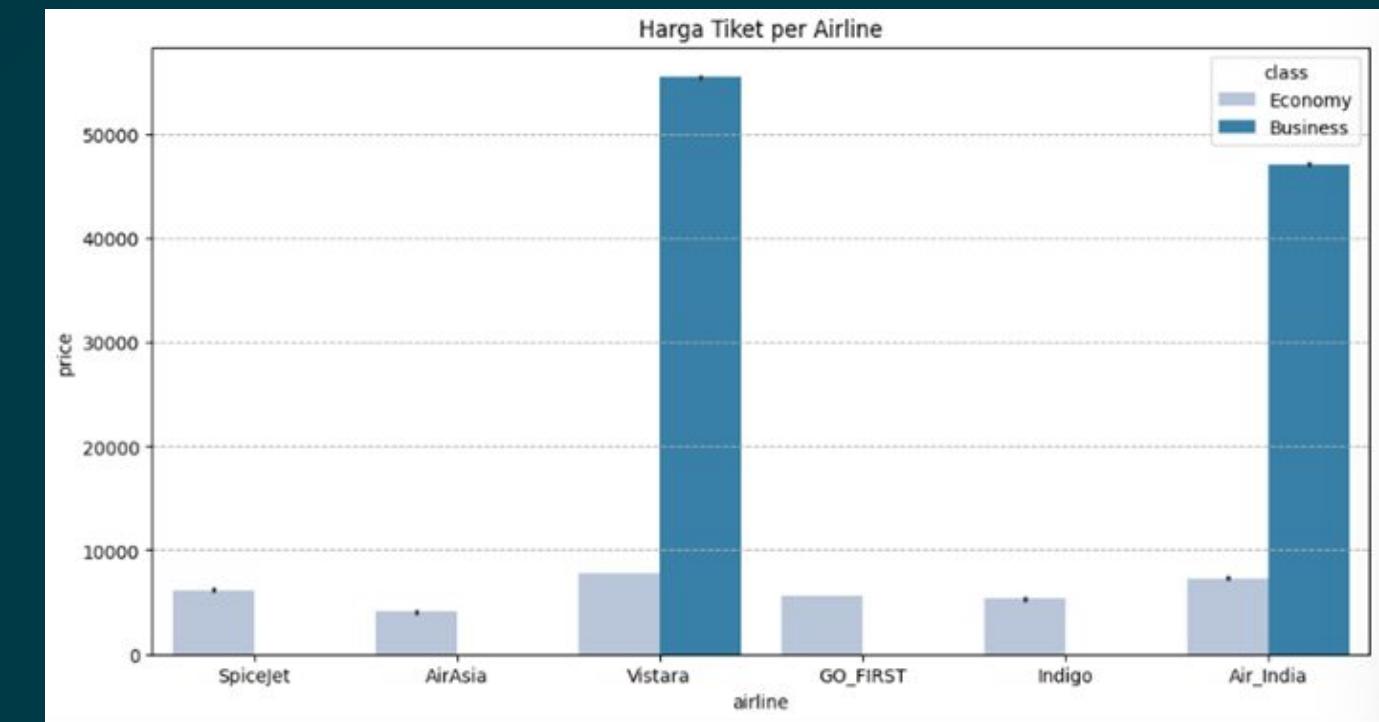
Gambar 4. Matrix Nilai Kosong atau Null

# HASIL DAN PEMBAHASAN

## VISUALISASI DATA



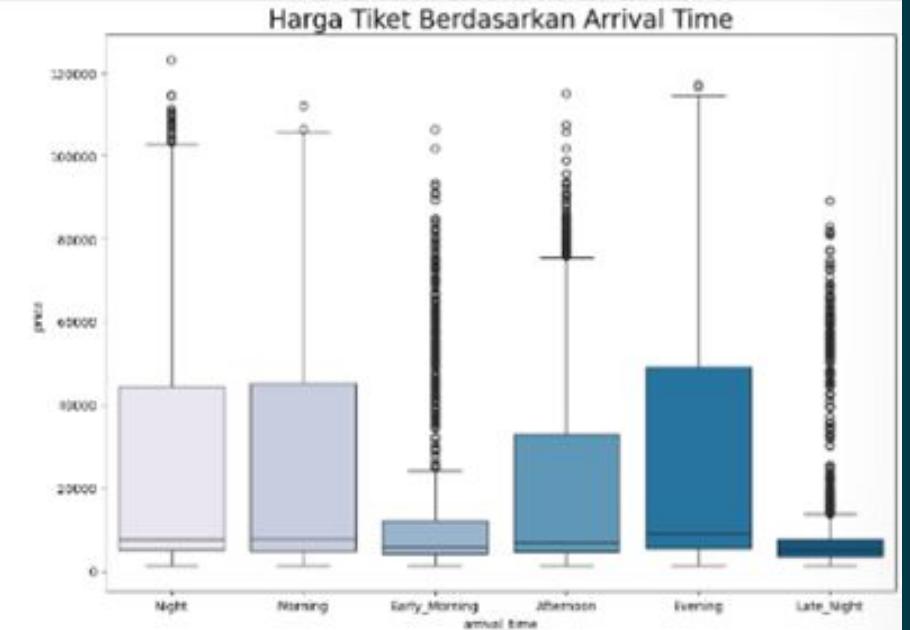
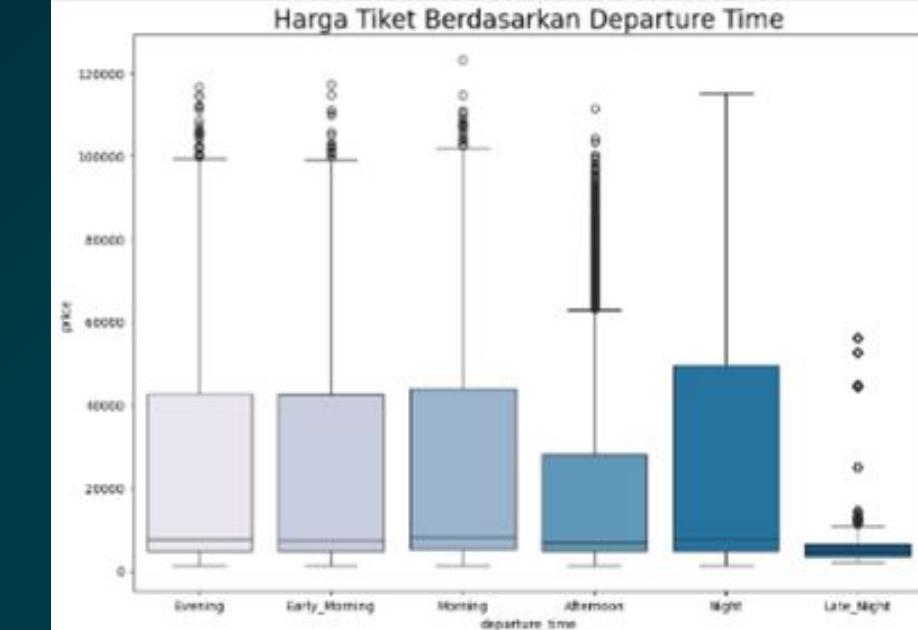
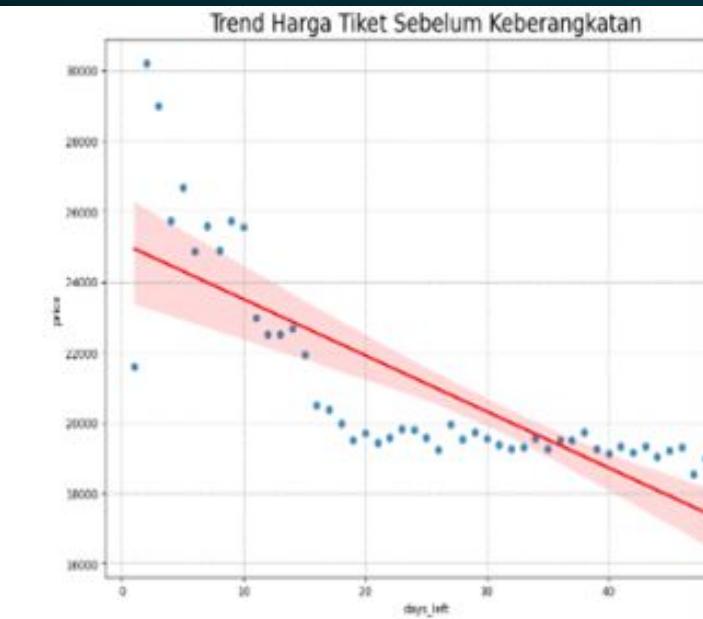
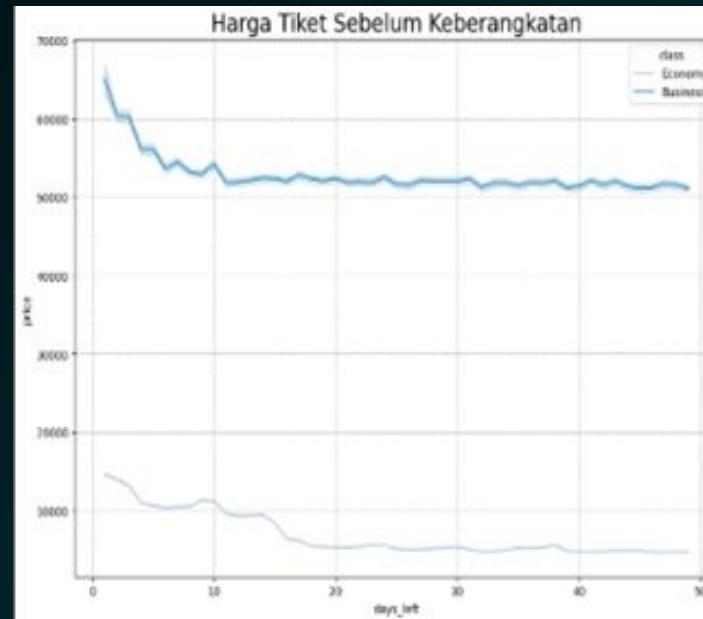
visualisasi data harga tiket pesawat pada kolom price yang dibandingkan dengan kolom class, airline, days\_left, departure\_time dan arrival\_time, serta sou



peningkatan harga tiket pesawat menjelang hari keberangkatan, akan dibuat visualisasi berdasarkan kolom price dengan days\_left

# HASIL DAN PEMBAHASAN

## VISUALISASI DATA

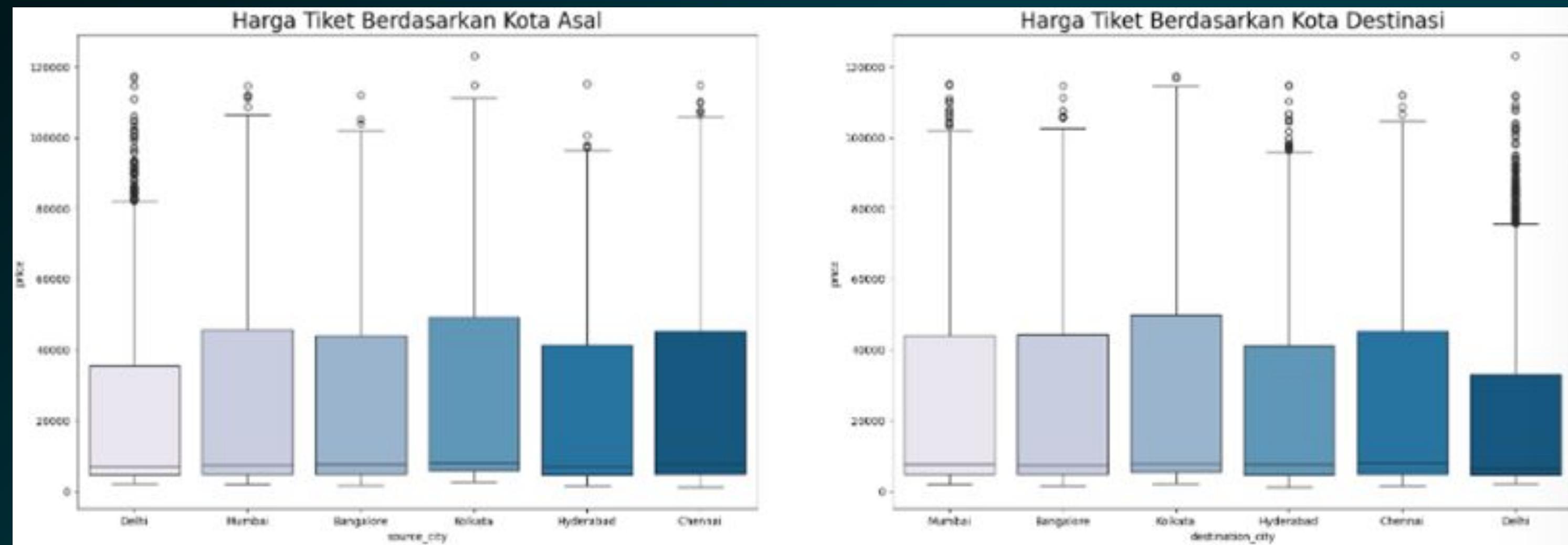


Tren Kenaikan Harga Tiket Sebelum Keberangkatan, dimana tren harga tiket akan semakin naik menjelang hari keberangkatan, terutama 20 hari menjelang keberangkatan harga tiket akan naik secara drastis.

Harga Tiket Berdasarkan Waktu Keberangkatan dan Kedatangan, visualisasi dilakukan pada kolom price dengan destination\_city dan source\_city

# HASIL DAN PEMBAHASAN

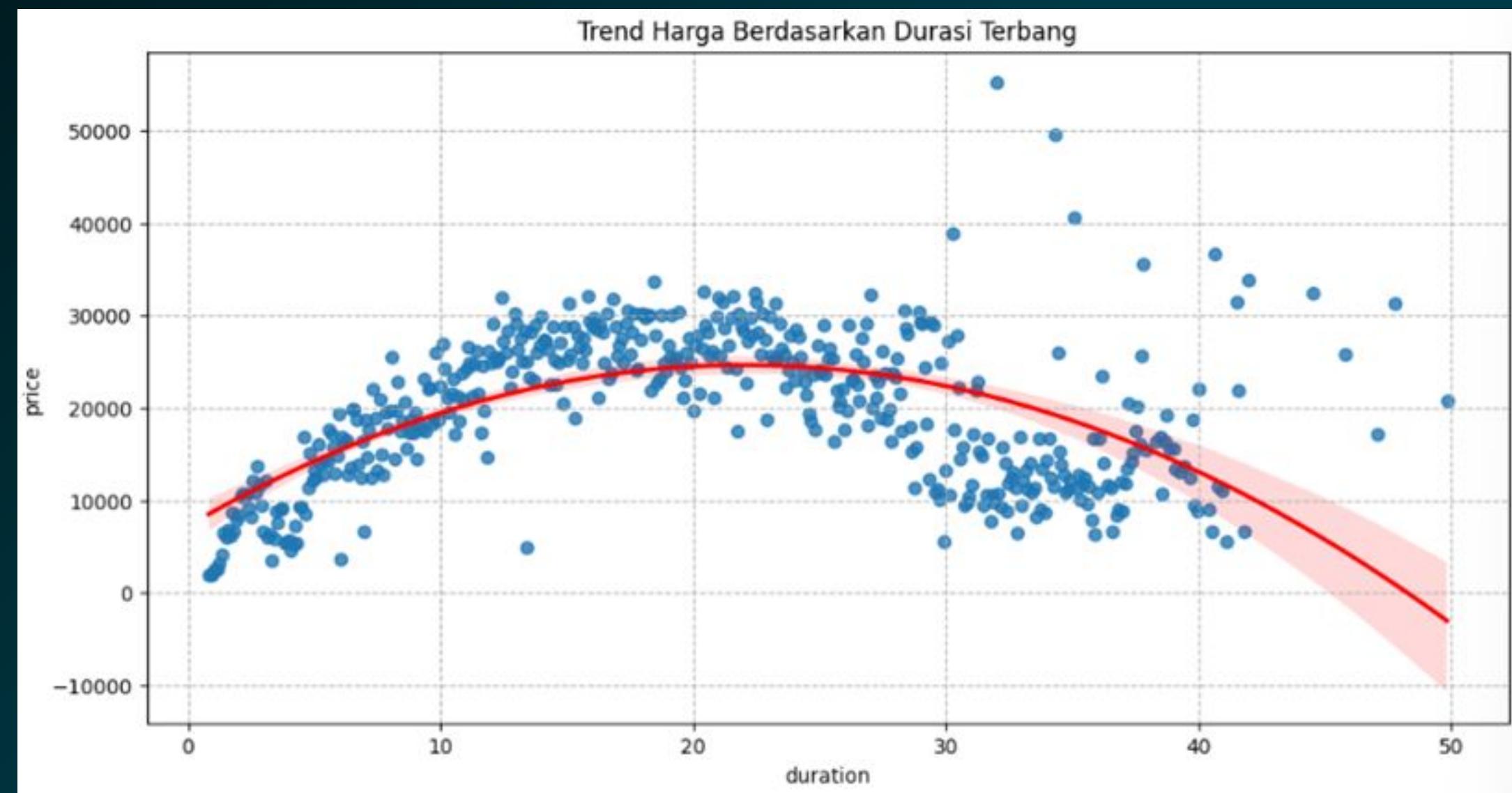
## VISUALISASI DATA



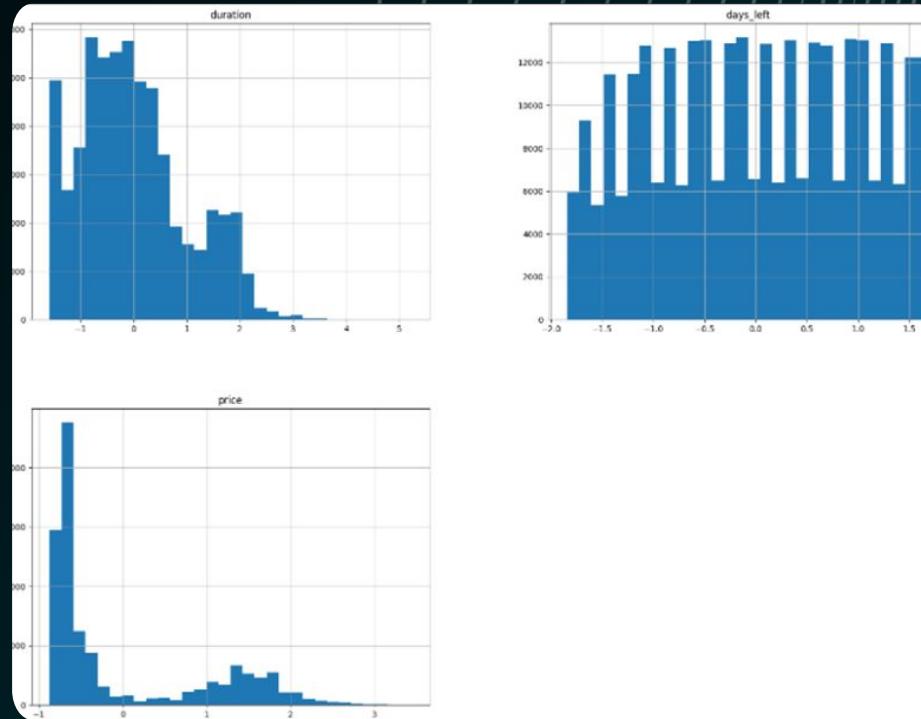
Harga Tiket Berdasarkan Kota Asal dan Kota Tujuan

# HASIL DAN PEMBAHASAN

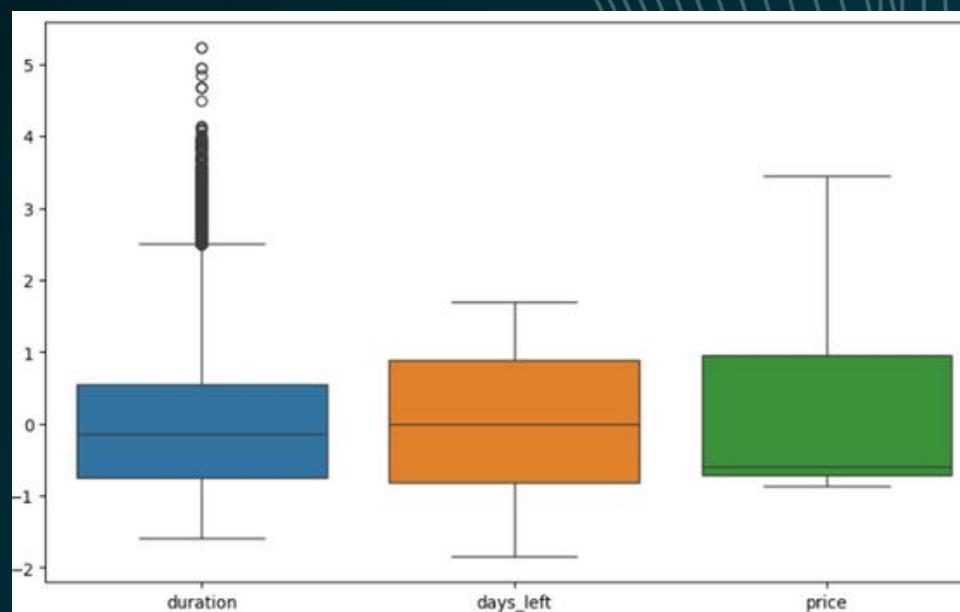
## VISUALISASI DATA



Tren Harga Tiket Berdasarkan Durasi Penerangan



Distribusi Data Fitur Numerik



Kanan : Distribusi Data Fitur Numerik

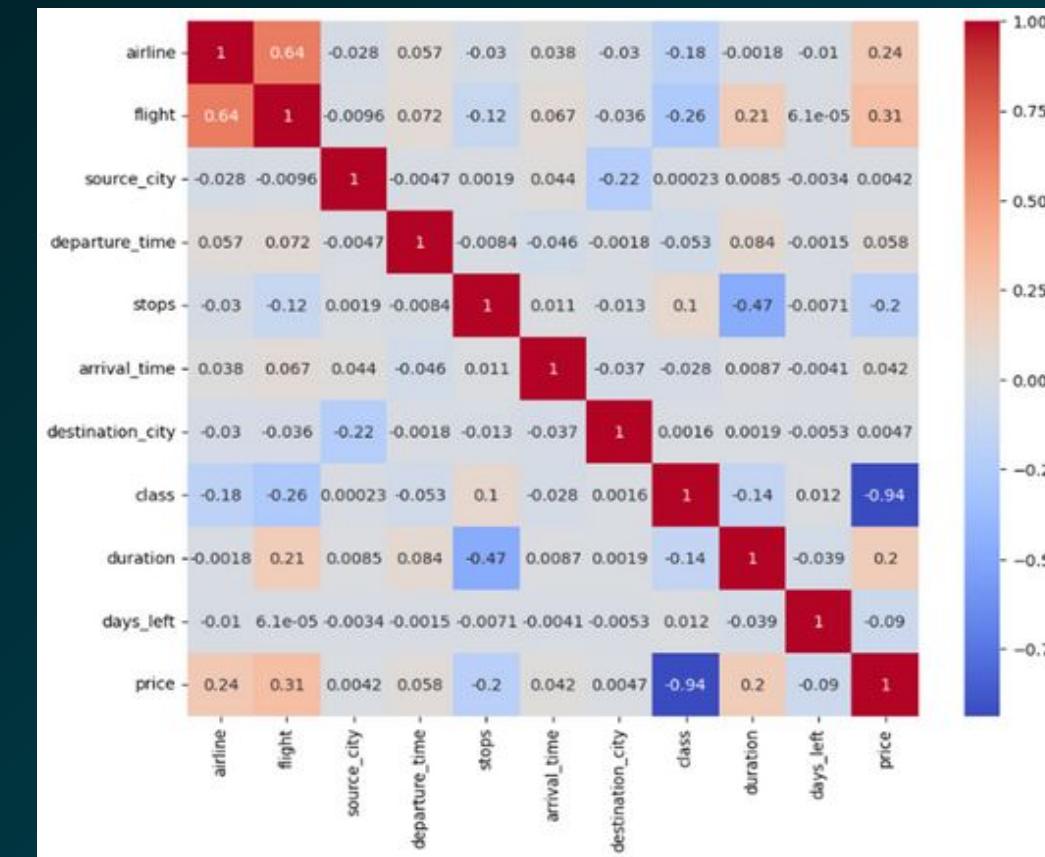
Atas : Persebaran Outlier dalam Box Plot

Setelah dilakukan Exploratory Data Analysis, selanjutnya adalah melakukan Preprocessing Data untuk mempersiapkan dataset yang akan dipakai untuk pemodelan. Langkah pertama adalah melakukan normalisasi data numerik dengan melakukan Scaling Data. Langkah selanjutnya adalah mengidentifikasi Outlier pada dataset. Sebagai contoh, dapat dilihat kembali pada gambar 8, 9 dan 10, dimana terdapat persebaran Outlier pada dataset.

Outlier dapat dihitung dengan mencari nilai Quartil 1 dan Quartil 3 dari kolom data yang dihitung, lalu selisihnya akan disebut sebagai Interquartile Range (IQR). Dan setelah dibersihkan, dapat dilihat pada gambar 11 informasi dari dataset yang sudah bersih dari Outlier dan pada gambar 12 merupakan nilai statistik dari 3 fitur numerik yang sudah dihapus Outlier dan dilakukan Scaling

dapat dilihat bagaimana visualisasi dari dataset yang sudah tidak memiliki Outlier lagi. Dimana persebaran data yang lebih seimbang pada 3 fitur numerik yang ada pada dataset. Serta pada gambar 16 merupakan bentuk dataset yang menampilkan 5 data teratas dari data yang sudah dilakukan penghapusan Outlier dan Scaling.

# HASIL DAN PEMBAHASAN



Correlation Matrix

Visualisasi matriks korelasi menggunakan heatmap membantu dalam membaca korelasi dengan memberikan warna yang mewakili kekuatan dan arah korelasi; warna yang lebih terang (seperti merah) menunjukkan korelasi positif yang kuat, sementara warna yang lebih gelap (seperti biru) menunjukkan korelasi negatif yang kuat. Dengan melihat visualisasi ini, kita dapat dengan cepat mengidentifikasi fitur mana yang berkorelasi kuat satu sama lain, yang penting untuk analisis fitur dan penanganan multikolinearitas.

# MODELING MACHINE LEARNING & DEEP LEARNING

	Neural Network	Random Forest	Linear Regression
MSE	0.6834800972271562	0.015749766069927847	0.0891925836485483
MAE	0.7212505568913203	0.06120027420536903	0.1985216213786529
R <sup>2</sup>	0.29416726694284456	0.983735151213257	0.9078904486897436

Hasil Evaluasi

Neural Network yang memiliki nilai MSE dan MAE cukup tinggi, serta nilai R<sup>2</sup> yang rendah menunjukkan bahwa model ini kurang akurat dan tidak mampu menjelaskan variabilitas data dengan baik. Pada Random Forest, nilai MSE dan MAE yang lebih rendah terutama MSE, serta R<sup>2</sup> yang hampir mendekati 1. Ini menunjukkan bahwa model Random Forest sangat akurat dan mampu menjelaskan hampir semua variabilitas data. Lalu terakhir Linear Regression memiliki nilai MSE dan MAE yang lebih baik dibanding Neural Network, terutama pada nilai MAE, serta R<sup>2</sup> yang cukup tinggi, tetapi masih dibawah Random Forest. Random Forest menunjukkan kinerja evaluasi terbaik karena beberapa faktor. Sebagai metode ensemble, ia menggabungkan prediksi dari banyak pohon keputusan, mengurangi variabilitas dan bias, serta mencegah overfitting dengan membangun pohon pada subset acak dari data.

# MODELING MACHINE LEARNING & DEEP LEARNING

feature	importance
class	0.881491
duration	0.048306
flight	0.028834
days_left	0.017297
destination_city	0.009447
source_city	0.005952
arrival_time	0.003189
departure_time	0.002637
stops	0.001919
airline	0.000927

Selanjutnya adalah melakukan optimasi pada model khususnya Random Forest. Dengan melihat fitur yang paling penting berdasarkan perangkingan, didapat hasil seperti pada gambar diatas, yaitu class sebagai fitur paling penting dan airline yang sangat tidak penting. Pada optimasi model Random Forest ini akan digunakan 2 metode Hyperparameter Tuning yaitu Randomized Search dan Grid Search.

# MODELING MACHINE LEARNING & DEEP LEARNING

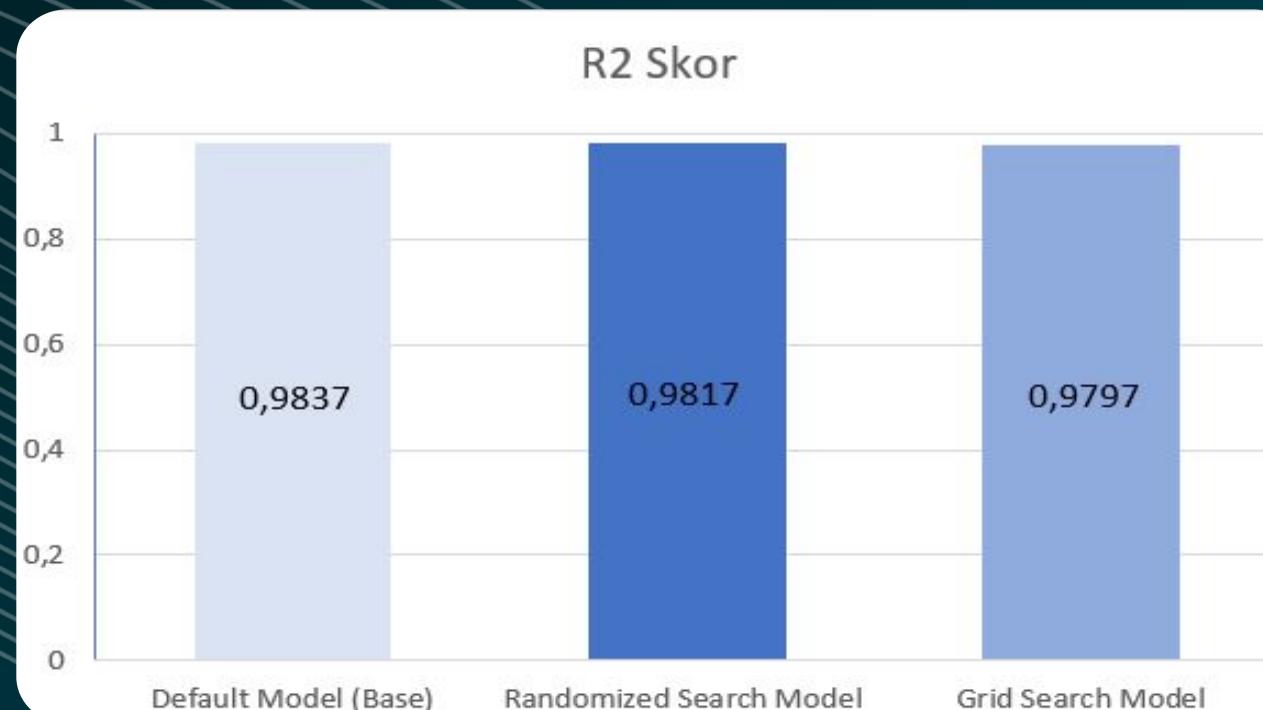
	<i>Randomized Search</i>	<i>Grid Search</i>
<b>Parameter Terbaik</b>	<pre>{'n_estimators': 350,  'min_samples_split': 5,  'min_samples_leaf': 2, 'max_features': 'sqrt',  'max_depth': 50, 'bootstrap': False}</pre>	<pre>{'bootstrap': True,  'max_depth': 40,  'max_features': 'sqrt',  'min_samples_leaf': 2,  'min_samples_split': 3, 'n_estimators': 300}</pre>
<b>Nilai Cross Validation Terbaik</b>	0.9812040585118588	0.9793021807758187
<b>MSE</b>	0.01794844673025553	0.019912706754155886
<b>MAE</b>	0.06842443635800396	0.075344220035221
<b>R<sup>2</sup></b>	0.9817257665758966	0.9797258527826946

Hasil Evaluasi pada Optimasi Model Random Forest

Hasil optimasi Random Forest menunjukkan bahwa kedua teknik hyperparameter tuning, Randomized Search dan Grid Search, memberikan hasil evaluasi yang sangat baik dengan beberapa perbedaan kecil. Meskipun kedua metode menghasilkan model yang sangat akurat, Randomized Search sedikit lebih unggul dalam semua metrik evaluasi, menunjukkan bahwa eksplorasi acak dari ruang parameter mungkin lebih efektif dalam menemukan kombinasi optimal dibandingkan dengan pencarian grid yang sistematis.

# KESIMPULAN

Penelitian ini menggunakan data dari "Ease My Trip" yang mencakup lebih dari 300.000 transaksi penerbangan antara enam kota metro utama di India. Setelah preprocessing, dataset dibagi menjadi data pelatihan dan uji. Dua model Machine Learning: Random Forest dan Linear Regression serta satu model Deep Learning Neural Network dilatih dan dievaluasi menggunakan metrik MSE, MAE, dan R2. Random Forest menunjukkan kinerja terbaik dengan nilai MSE dan MAE terendah serta R2 tertinggi, sementara Neural Network menunjukkan performa terendah. Optimasi model Random Forest dilakukan melalui Randomized Search dan Grid Search, menghasilkan model yang lebih robust, akurat dan generalis meskipun terdapat sedikit penurunan pada nilai R2 dibandingkan model default, menunjukkan bahwa tuning meningkatkan ketahanan model. Penelitian ini mengonfirmasi bahwa Random Forest adalah model terbaik untuk memprediksi harga tiket pesawat.



Perbandingan Skor R2 pada Model Random Forest

pada penelitian ini juga memberikan wawasan penting bagi berbagai pemangku kepentingan, termasuk konsumen, maskapai penerbangan, dan pembuat kebijakan. Konsumen dapat memanfaatkan informasi mengenai variasi harga antar maskapai dan waktu optimal pembelian untuk menghemat biaya.



FINAL PROJECT DATA SCIENCE & ANALYST

# TERIMA KASIH



Disusun oleh Kelompok 5

