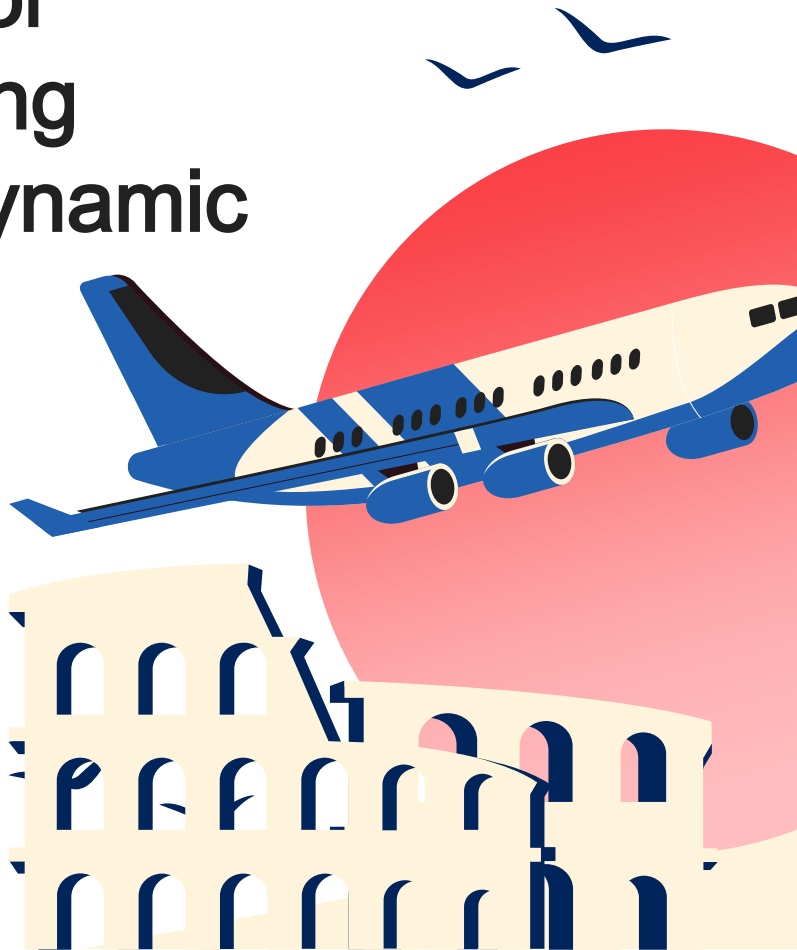
Utilizing Big Data Tools for Agoda's Customer Booking Behavior Analysis and Dynamic Pricing Optimization

WQD7007 Occ1 – Group 2 & 8

LECTURER: TS. DR. MOHD SHAHRUL NIZAM BIN MOHD DANURI

Name	Student ID	Tools
Nur Hidayah binti Ahmad Shafii (Leader)	22120931	MapReduce
Diva Alifta Chandra	23069683	MapReduce
Yong Ting Kang	23083416	Hive
Zeng Yan Ting	S2168467	Spark
Boaz Chung Yi Heng	23059592	Spark
Then Dao Qing	23057608	Hbase + PowerBI
Choon Yue Hua	17152027	HDFS + PowerBI
Loh Bi Jia	23078886	Apache Pig + Python



INTRODUCTION

Agoda was founded in 2005 by Michael Kenny and Robert Rosenstein in Phuket, Thailand.

Diverse customer purchasing behaviors and determining accurate pricing remains a challenge. Risk of losing competitive edge, revenue opportunities, and customer satisfaction without effective solutions.

OBJECTIVES

This study aims to evaluate the effectiveness of Hadoop in improving Agoda's ability to achieve the following objectives:

1. To analyse booking behaviours based on demographic and booking history
2. To identify patterns that can inform dynamic pricing strategies

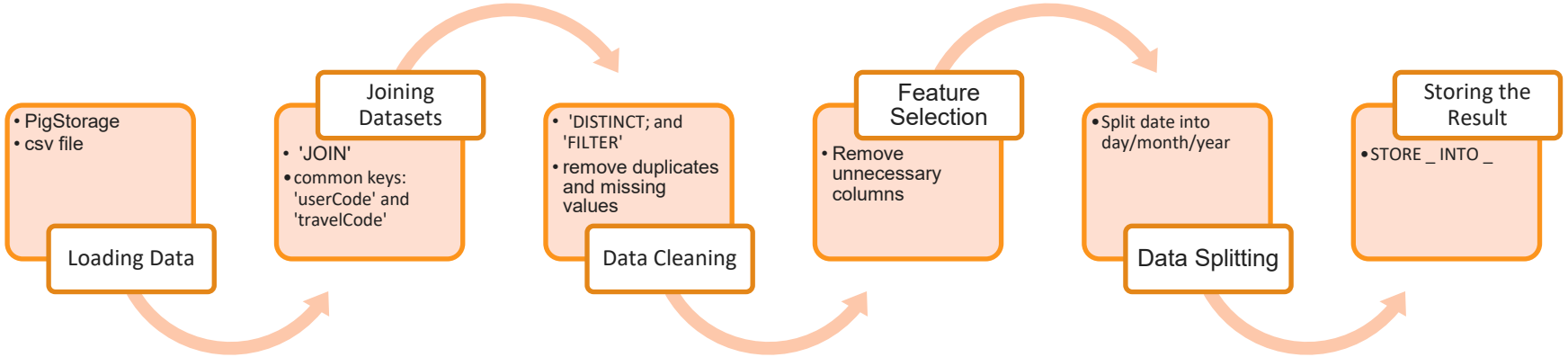
TOOLS/TECHNOLOGIES

Apache Pig/Apache Hive/ MapReduce/ Spark/HDFS/ Hbase/ PowerBI/ Python

01

Data Preprocessing: Apache Pig and Python

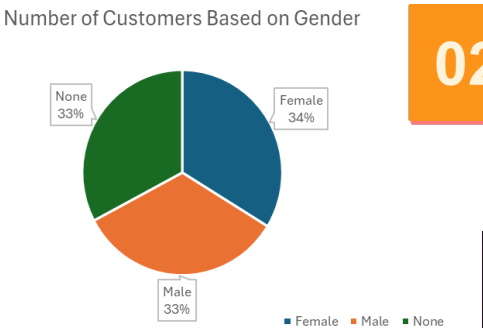
Apache Pig



Python is used to enrich the datatset.
The day of the week for the flight and hotel dates is done through Python.

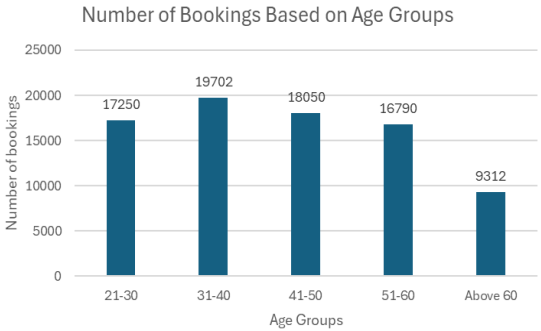
02 Data Querying (Apache Hive & MapReduce)

Customer Demographics



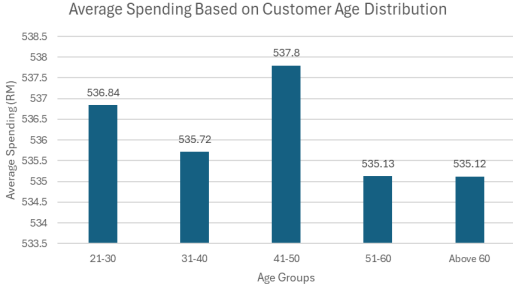

```
OK
female  27380
male    27170
none    26554
Time taken: 47.238 seconds,
```

```
OK
21-30   17250
31-40   19702
41-50   18050
51-60   16790
Above 60        9312
Time taken: 58.415 seconds, F
```



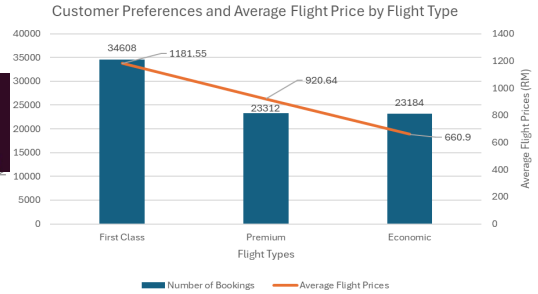
Pie chart shows a relatively **balanced proportion** of customers among the three gender categories. Agoda should cater marketing strategies to a **diverse range of customers.**

Bar chart illustrates the age group of 31-40 is the most active demographic which followed by age group of 41-50 and 21-30. Agoda can target on **middle-aged and young adults groups** to tailor the travel packages and services.




```
OK
21-30   17250
31-40   19702
41-50   18050
51-60   16790
Above 60        9312
Time taken: 58.415 seconds, Fetched: 5 row(s)
```

```
OK
firstClass      34608   1181.55
premium 23312   920.64
economic        23184   660.9
Time taken: 55.24 seconds, Fetched: 3 row(s)
```



Bar chart displays the average spending analysis across different age groups which the **average range** of **RM 535–RM 538**. The highest average spending is observed in the **41-50 age group** at RM 537.80, followed by age group of 21-30 and 31-40. This pattern suggests **similar purchasing behavior** among these demographics.

Based on the bar charts, economic class showed the most cost-effective option with average price of RM 660.90. However, majority of customers prefer with first class (34,608), premium class (23,312) and economic class (23,184). This pattern reflects **customer prefer with exclusivity and superior service** on flight selection.

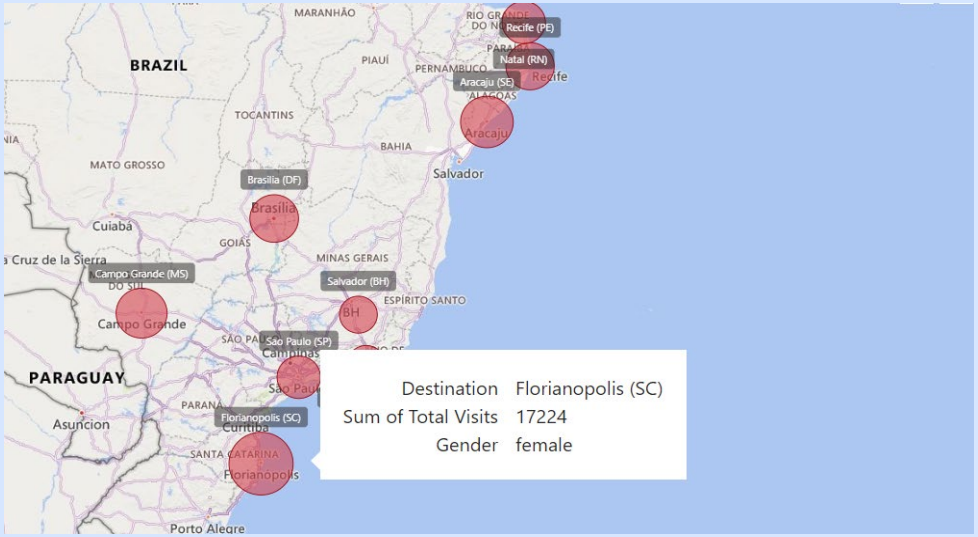
Customer Behavior

MapReduce

```
vboxuser@ubuntu1:~$ cat /home/vboxuser/clea
female  Florianopolis (SC)      5816
female  Aracaju (SE)    3760
female  Campo Grande (MS)       3302
female  Brasilia (DF)   3137
female  Recife (PE)     3048
female  Sao Paulo (SP)  2459
female  Natal (RN)      2411
female  Salvador (BH)   1733
female  Rio de Janeiro (RJ)     1714
gender  to      1
male    Florianopolis (SC)      6037
male    Aracaju (SE)    3631
male    Campo Grande (MS)       3334
male    Brasilia (DF)   2925
male    Recife (PE)     2908
male    Natal (RN)      2488
male    Sao Paulo (SP)  2409
male    Salvador (BH)   1759
male    Rio de Janeiro (RJ)     1679
none    Florianopolis (SC)      5371
none    Aracaju (SE)    3665
none    Campo Grande (MS)       3635
none    Brasilia (DF)   3166
none    Recife (PE)     3123
none    Sao Paulo (SP)  2202
none    Natal (RN)      2154
none    Rio de Janeiro (RJ)     1636
none    Salvador (BH)   1602
vboxuser@ubuntu1:~$
```

Apache Hive

```
Total MapReduce CPU Time Spent: 12 seconds 620 m
+---------+----------------------+-------------+
| gender  |     destination      | total_visits |
+---------+----------------------+-------------+
| female  | Florianopolis (SC)   | 5816        |
| female  | Aracaju (SE)         | 3760        |
| female  | Campo Grande (MS)    | 3302        |
| female  | Brasilia (DF)        | 3137        |
| female  | Recife (PE)          | 3048        |
| female  | Sao Paulo (SP)       | 2459        |
| female  | Natal (RN)           | 2411        |
| female  | Salvador (BH)        | 1733        |
| female  | Rio de Janeiro (RJ)  | 1714        |
| gender  | to                   | 1           |
| male    | Florianopolis (SC)   | 6037        |
| male    | Aracaju (SE)         | 3631        |
| male    | Campo Grande (MS)    | 3334        |
| male    | Brasilia (DF)        | 2925        |
| male    | Recife (PE)          | 2908        |
| male    | Natal (RN)           | 2488        |
| male    | Sao Paulo (SP)       | 2409        |
| male    | Salvador (BH)        | 1759        |
| male    | Rio de Janeiro (RJ)  | 1679        |
| none    | Florianopolis (SC)   | 5371        |
| none    | Aracaju (SE)         | 3665        |
| none    | Campo Grande (MS)    | 3635        |
| none    | Brasilia (DF)        | 3166        |
| none    | Recife (PE)          | 3123        |
| none    | Sao Paulo (SP)       | 2202        |
| none    | Natal (RN)           | 2154        |
| none    | Rio de Janeiro (RJ)  | 1636        |
| none    | Salvador (BH)        | 1602        |
+---------+----------------------+-------------+
28 rows selected (89.461 seconds)
0: jdbc:hive2://>
```



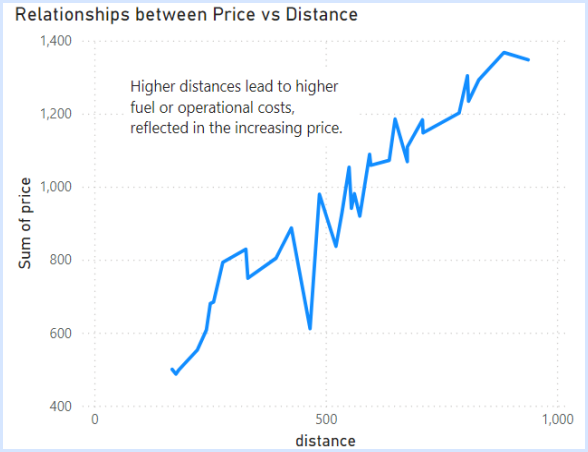
The geomap highlights Florianopolis (SC) as one of the most popular destinations among visitors. Specifically, female visitors accounted for 5,816 total visits. This indicates a significant preference among female travelers compared to other destinations.

Customer Behavior

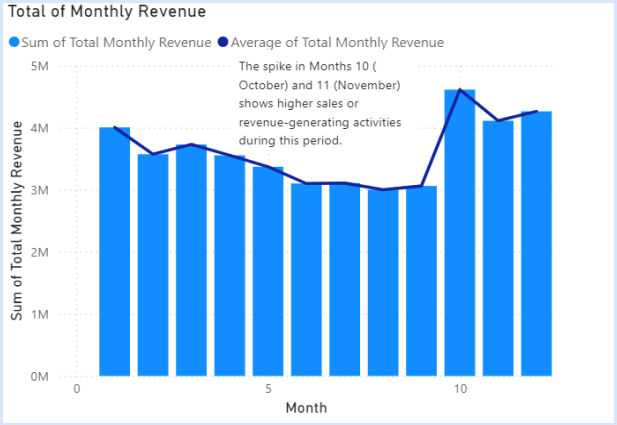
Customers with Highest Revenue

Full Name	Total Revenue
Helen Warner	81,999.72
Wallace Gallardo	78,809.28
Ray Johnson	78,376.22
Andrew Anderson	78,130.92
John Micciche	78,004.30
Kevin Paul	77,977.16
Linda Ellis	76,823.52
Juanita Palmer	76,493.84
Kenneth Jump	75,607.94

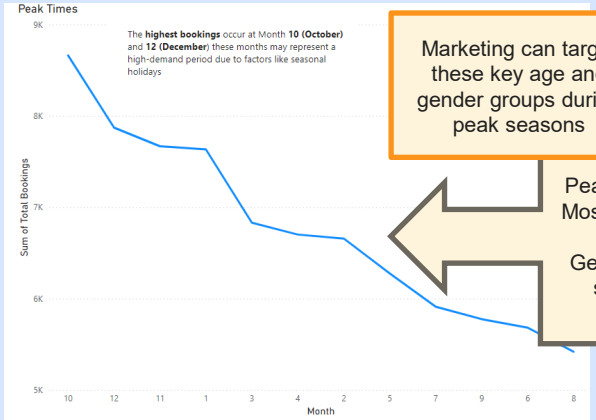
Top customer: Helen Warner ($81,999.72)
Top 8 Customers contribute to total revenue



Price increases with greater distances due to higher fuel/operational costs

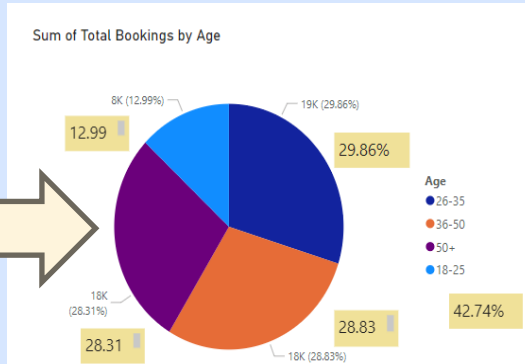


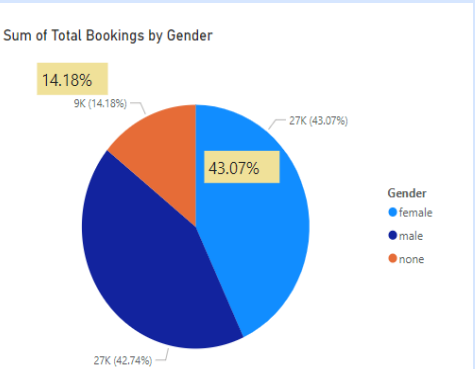
October and November generate the highest monthly revenue, highlighting a peak season for customer activity



Marketing can target these key age and gender groups during peak seasons

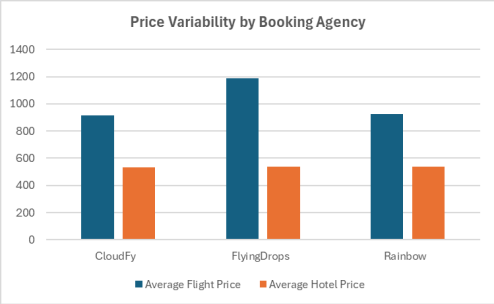
Peak Times : Oct & Nov
Most Active Groups : 30-35 and 25-30
Gender : Female leads slightly over Male





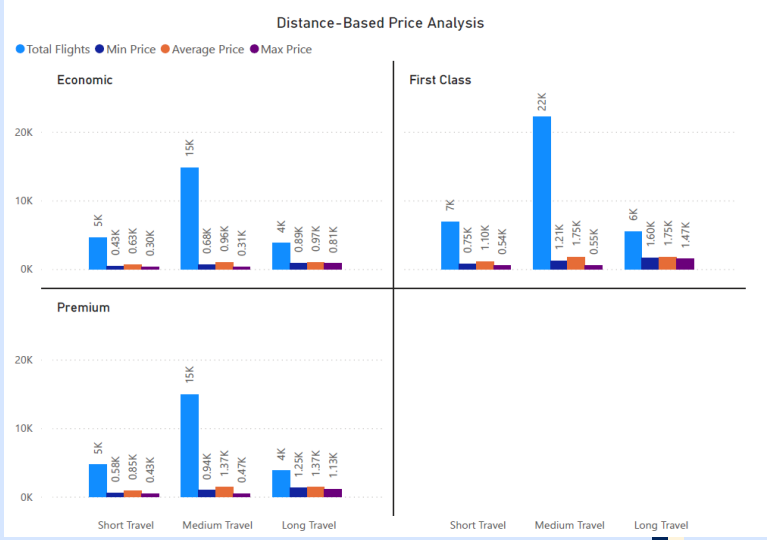
Price Optimization

Salvador: Highest average flight cost (RM1,179.23) & hotel prices (RM263.41).
Sao Paulo: Lowest flight costs (RM826.55) & affordable hotel prices (RM139.10).



Highest Flight Prices: FlyingDrops (RM 1,186.13)
Competitive Flight Pricing: Rainbow (RM 922.96) & CloudFy (RM 917.02).
Stable Hotel Prices: Minimal differences across agencies (RM 534.99–RM 539.72).

Flight numbers declined from January to August (lowest: 5420 flights).
Peak travel in October (8663 flights) followed by December and November, driven by holiday demand.



Short Travel (≤300 km):
- Economic: Most cost-effective (RM 629.54, 4,628 flights).
- Premium: 35.2% pricier than Economic

Medium Travel (301–800 km):
- Most in-demand for all classes

Long Travel (>800 km):
- Reduced demand across all classes.

Comparison Performance between Hive and MapReduce Tools

Tools	Hive	MapReduce
Stages/Jobs	2 Stages	2 Jobs
Mappers	1 mapper per stage	Job 1: 2 mappers; Job 2: 2 mappers
Reducers	1 reducer per stage	1 reducer per job
Cumulative CPU Time	Stage-1: 9.13 sec; Stage-2: 5.77 sec	Job 1: 7.06 sec; Job 2: 14.90 sec
HDFS Read	Stage-1: 13,467,419 bytes; Stage-2: 8,296 bytes	13,451,526 bytes
HDFS Write	Stage-1: 529 bytes; Stage-2: 458 bytes	263 bytes
Input Records	Not explicitly mentioned	Job 1: 81,105
Output Records	Not explicitly mentioned	Job 1: 81,104; Job 2: 9
Memory Usage (Peak)	Not explicitly mentioned	Map: ~291 MB; Reduce: ~182 MB
Execution Time	Stage-1: ~18 sec; Stage-2: ~16 sec	Job 1: ~27 sec; Job 2: ~33 sec

Performance Insights:

- Hive is ideal for analysis-focused tasks with smaller to medium datasets.
- MapReduce excels in handling large datasets and custom processing.

Recommendation:

- Use Hive for this project due to simplicity, speed, and practicality.
- Consider MapReduce for scalability if the dataset grows.

Hive vs. MapReduce:

- **Hive:** User-friendly, SQL-like interface; faster execution (~14.9 seconds), efficient memory management, and lower disk usage.
- **MapReduce:** Offers detailed control; slower execution (~27–33 seconds), higher data transfer, and manual memory configuration (~291 MB map, ~182 MB reduce).

03

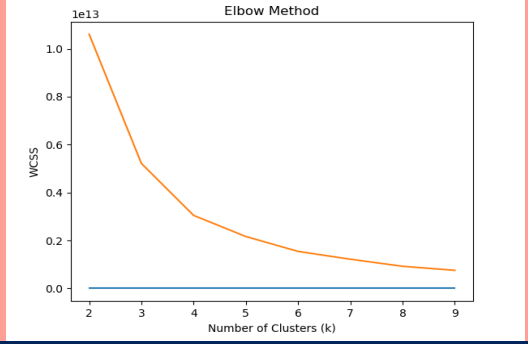
HDFS - Data Storage and Integration

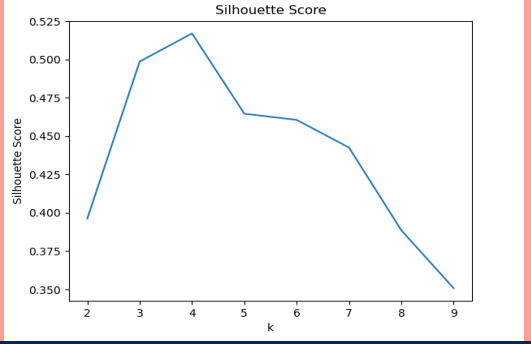
Centralized Data Storage	Data Partitioning and Replication	Seamless Integration with Processing Tools
Acts as the main repository for raw, intermediate, and processed data.	Automatically divides datasets into smaller blocks.	Supports efficient data processing.
	Replicates data across nodes for high availability and fault tolerance.	**Pig:** Preprocessing raw data. **Hive:** Querying processed data. **MapReduce:** Advanced analysis for insights. **Spark:** Model training and evaluation.

04

K-Means Clustering: Apache Spark

Cluster Initialization





Optimal k=4

Cluster Center Coordinates

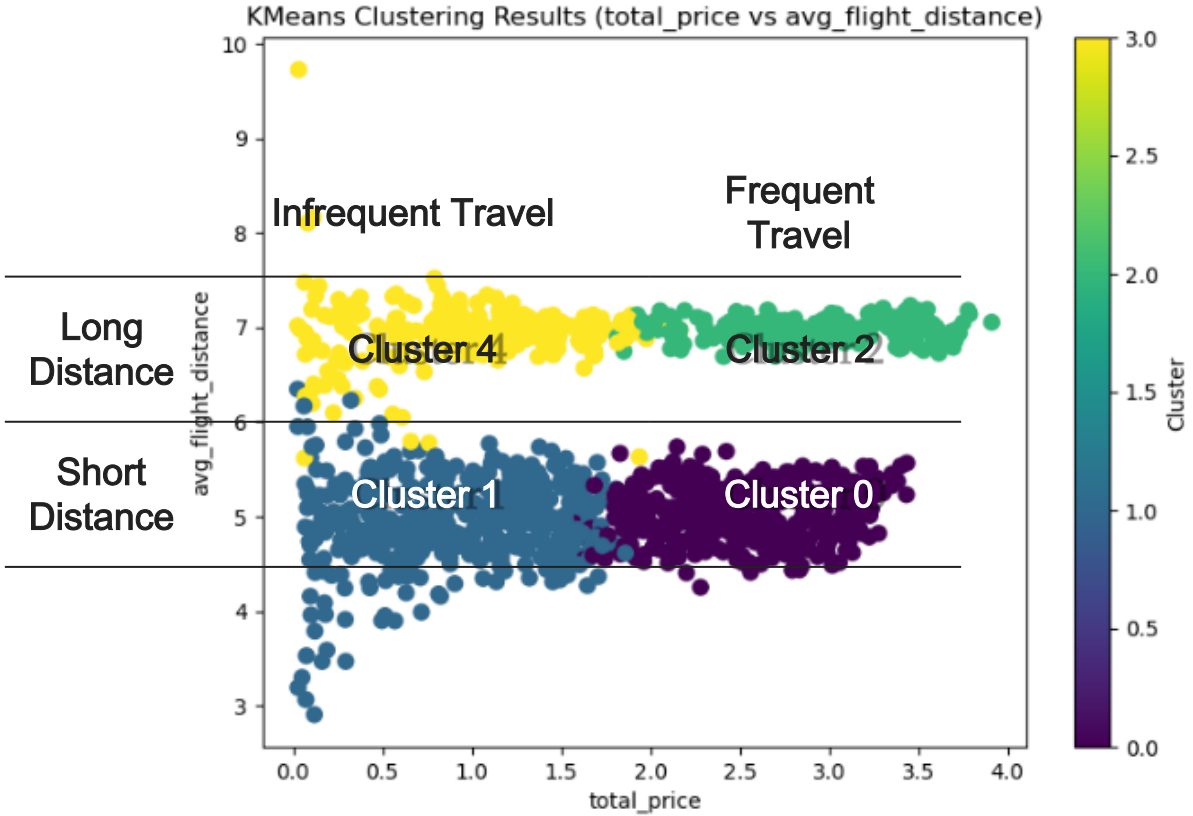
% Difference from average position

cluster	total_ mileage	total_ flight_ price	total_ flights	total_ days_ hotel	total_ hotel_ price	total_ hotels	total_ price	avg_ flight_ distance	avg_ flight_ time	avg_ flight_ price	avg_ hotel_ price_ daily	age
0	23.9%	36.8%	47.4%	47.6%	53.2%	47.6%	38.0%	-15.8%	-15.8%	-7.1%	4.3%	1.9%
1	-55.3%	-50.4%	-46.0%	-47.1%	-45.7%	-47.2%	-50.1%	-17.2%	-17.2%	-8.3%	3.4%	0.6%
2	72.7%	59.4%	48.1%	50.7%	45.4%	50.7%	58.4%	16.8%	16.7%	7.7%	-3.2%	-1.8%
3	-41.3%	-45.8%	-49.4%	-51.2%	-53.0%	-51.1%	-46.3%	16.2%	16.2%	7.7%	-4.5%	-0.7%

Frequency of travel

Distance

Visualization of Clusters

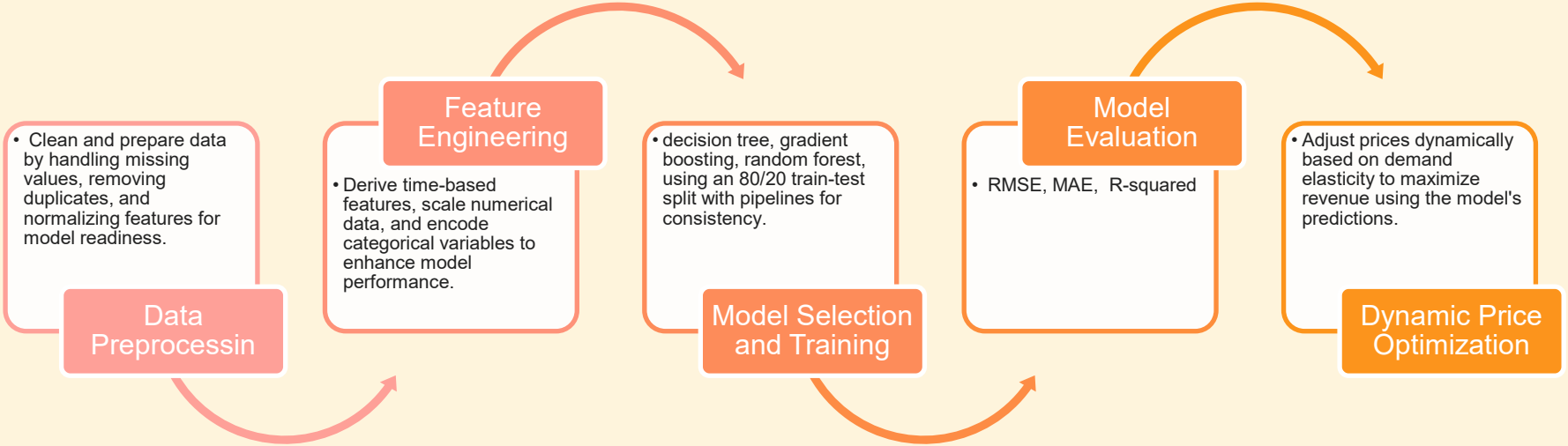


05

Modeling Training & Evaluation

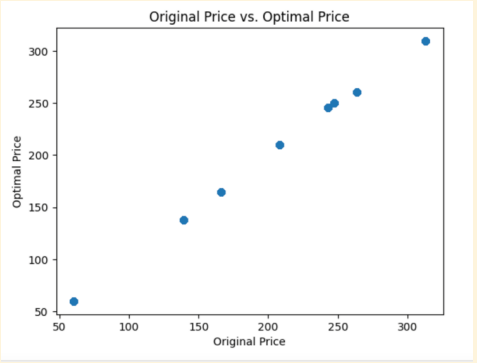
Regression Modeling

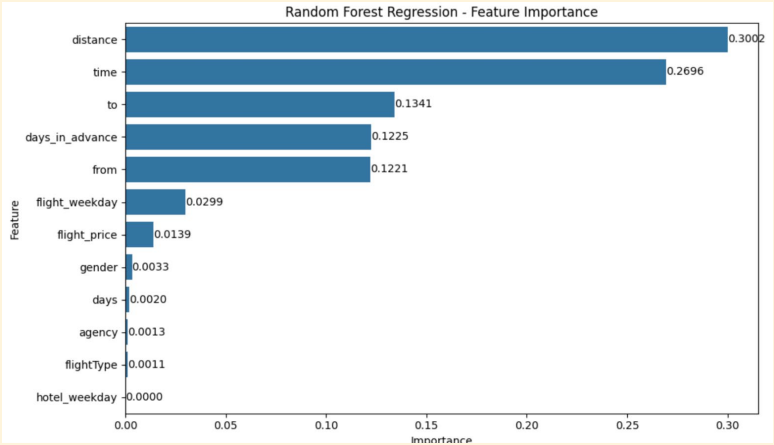
The goal of this stage is to adjust the hotel price (hotel_price) to optimise revenue based on the prediction results of the best regression model. This study aims to dynamically adjust the hotel price to maximise revenue based on the prediction results of the best regression model.

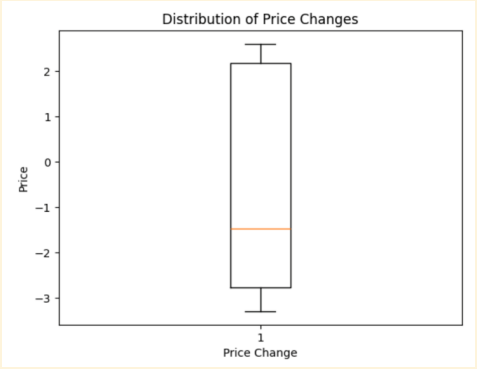


Regression Model Evaluation

Model Name	RMSE	MAE	R-squared
Random Forest Regression	20.2	12.31	0.931
Gradient Boosting Regression	20.49	14.02	0.929
Decision Tree Regression	28.3	8.15	0.865







06 Visualization

A

Extract and Prepare

- HBase data extraction
- Aggregate booking metrics (counts, price)
- Transform for visualization

B

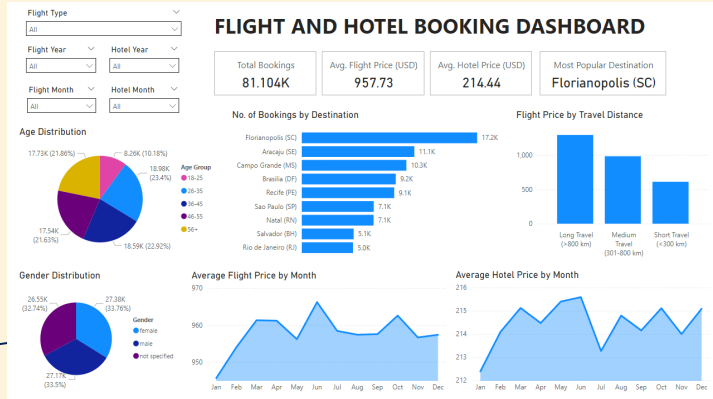
Implement and Design

- Model data relationships in Power BI
- Create intuitive visualizations

C

Validate and Deploy

- Test dashboard performance
- Verify data accuracy



CONCLUSION & RECOMMENDATION

Empowering Targeted Marketing:
- Focus on middle-aged travelers (31–50 years).
- Design campaigns tailored for peak seasons (October, December) and off-peak opportunities (August).

Optimizing Pricing Strategies:
- Use dynamic pricing to balance peak demand and off-peak promotions.
- Introduce premium packages for high-end destinations like Salvador and Natal.

Promoting Personalized Offers:
- Use K-Means clustering insights for targeted promotions based on travel frequency and distance.

Enhancing Technological Capabilities:
- Utilize Hive for efficiency (use MapReduce if the datasets grow)
- Explore advanced machine learning for real-time insights and dynamic decision-making.