

US Airline Passenger Satisfaction

Business Intelligence, May 2022



Our Team



Bruno Faria

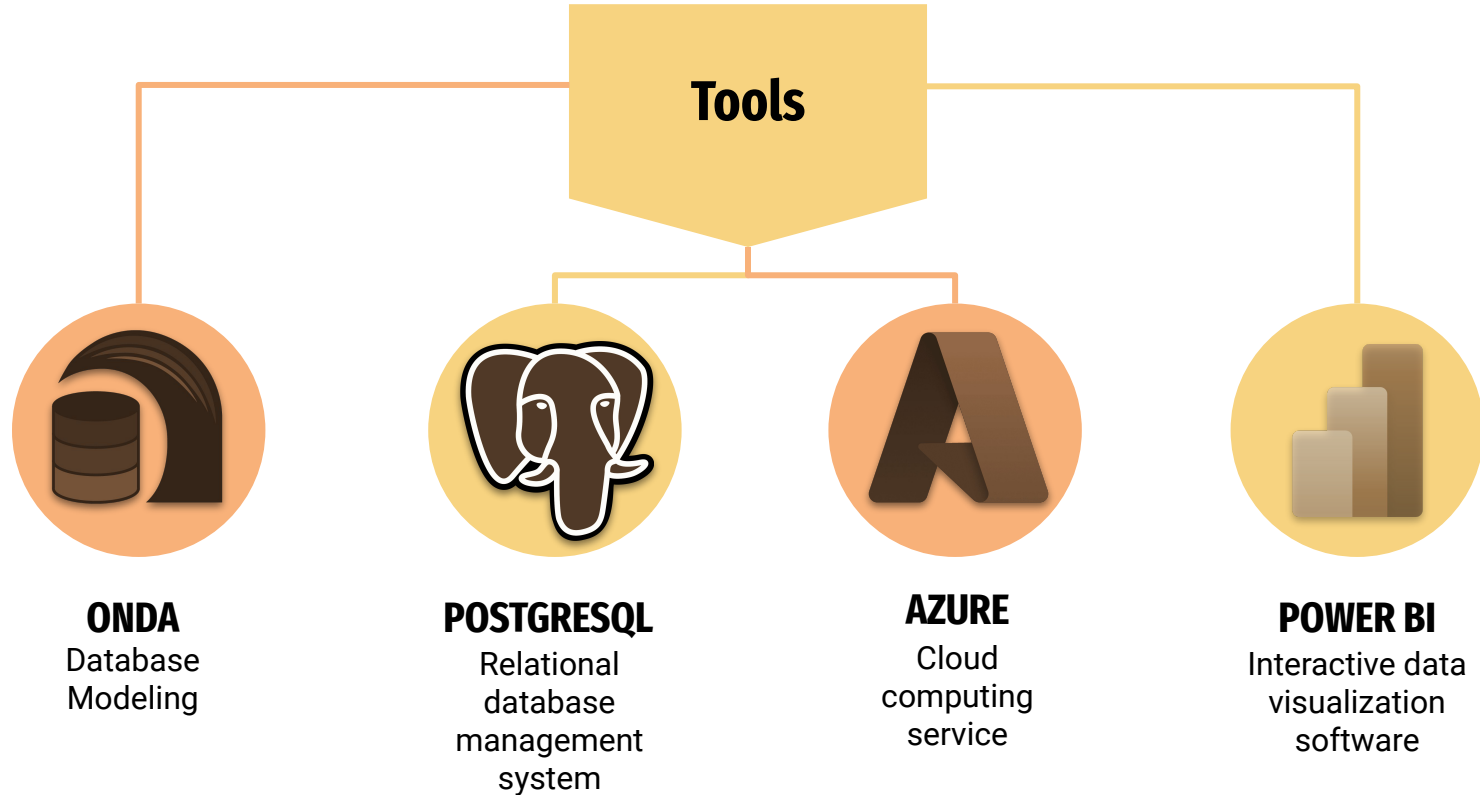
2018295474
brunofaria@student.dei.uc.pt



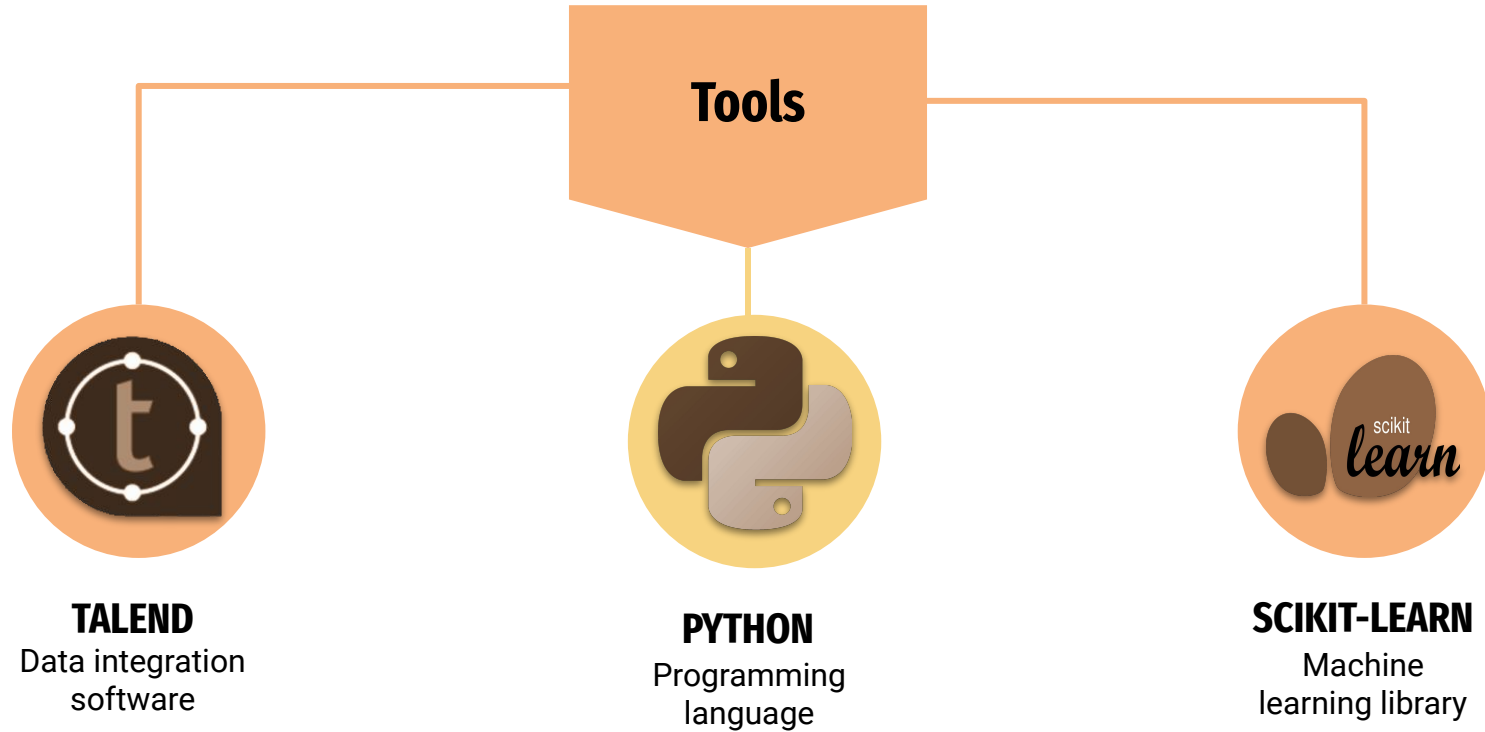
Dylan Perdigão

2018233092
dgp@student.dei.uc.pt

Tools



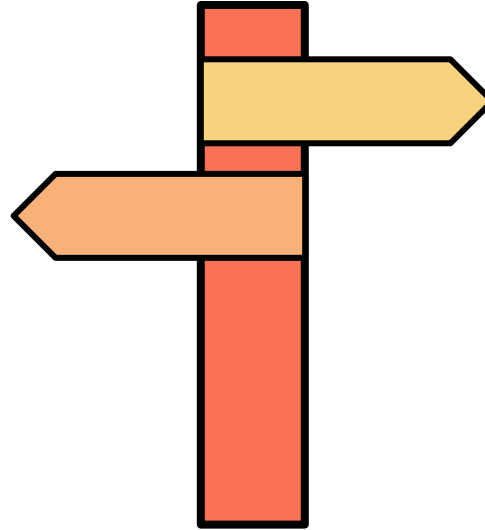
Tools



Goals

**What's the average classification
for the provided services?**

**Which kind of passenger
travels the most?**



**Who are the most
satisfied customers?**

**What's the Age and Distance
travelled distribution?**

Predict the satisfaction using machine learning techniques

Dataset



Provenance

Kaggle (see [here](#))

Rows

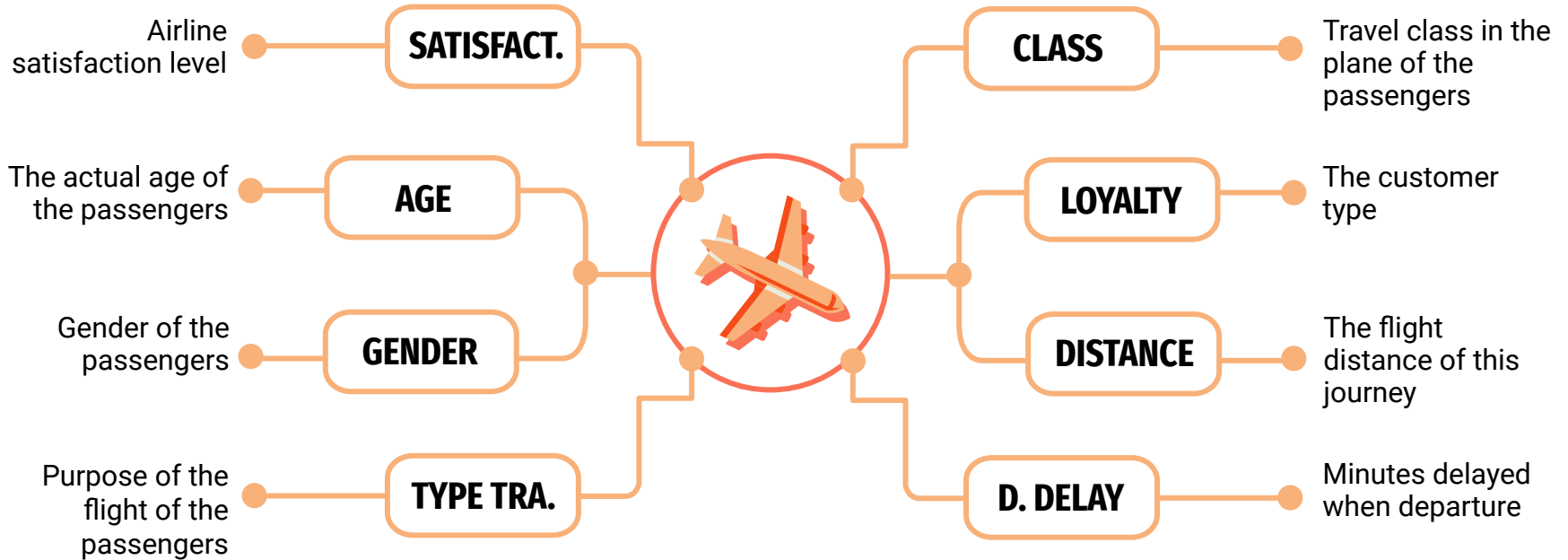
129880 instances

Columns

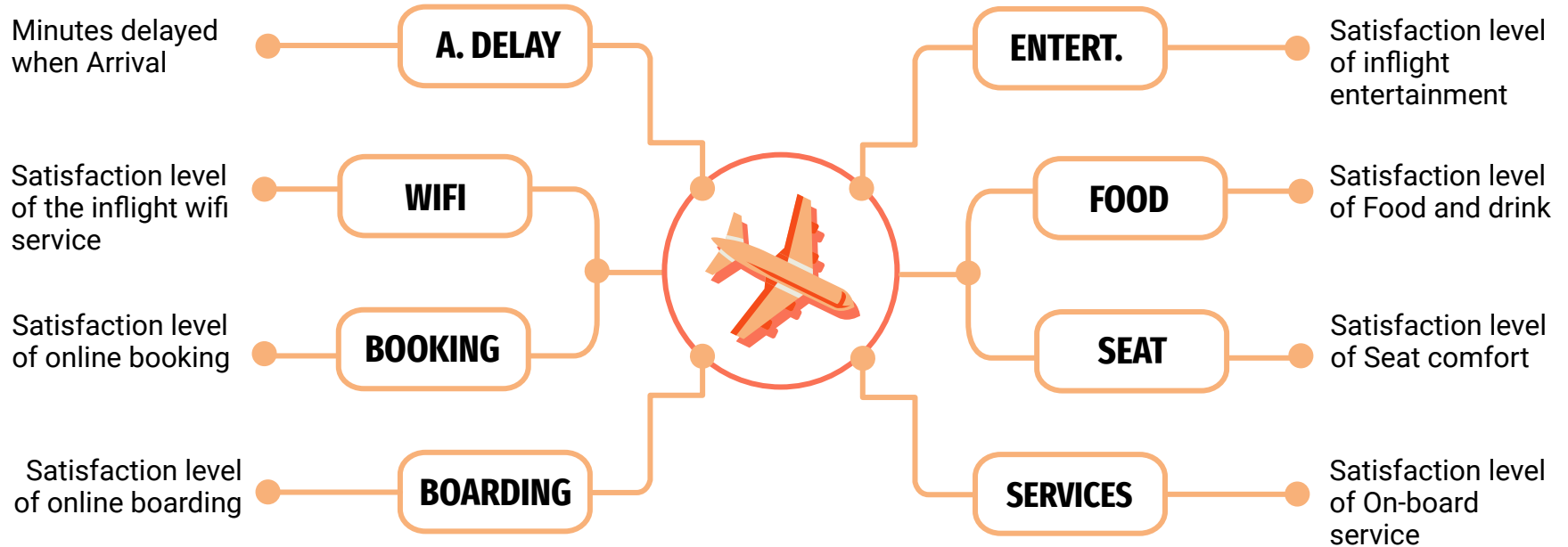
23 columns



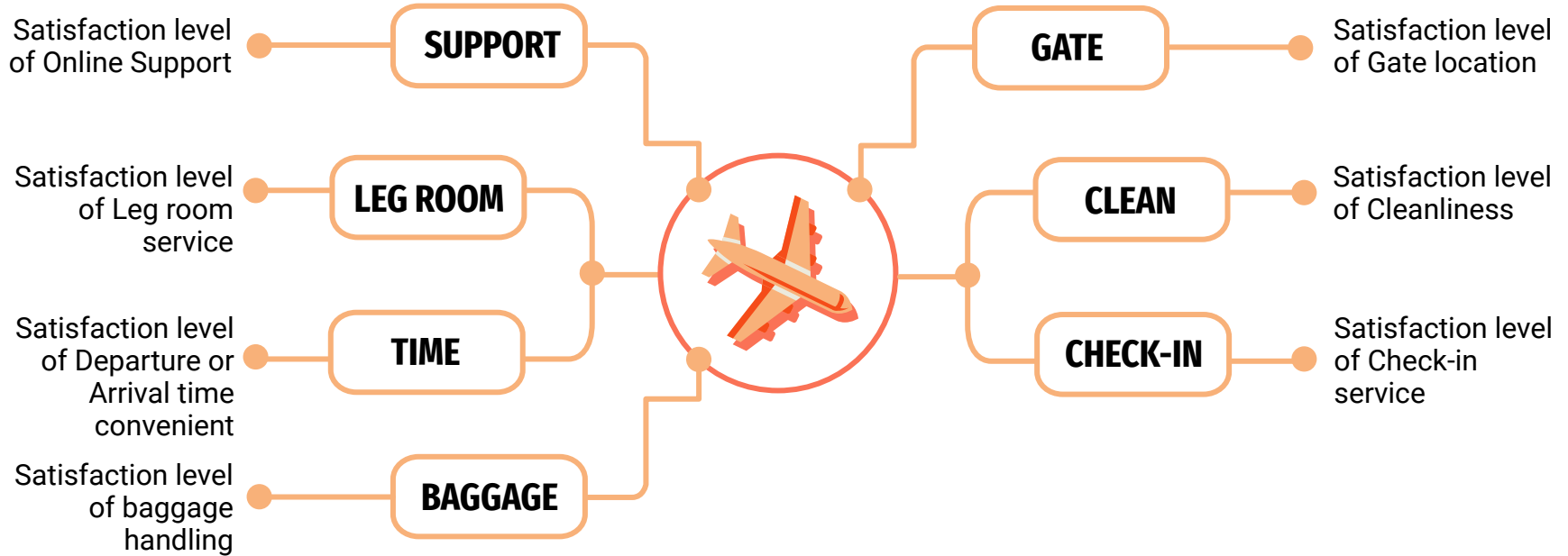
Dataset



Dataset



Dataset



Data Design

PERSON DIMENSION

Passenger basic
personal data

Person			
id_person	BInt	PK	
age	Int	NN	
gender	VChr	NN	
loyalty	VChr	NN	

SATISFACTION DIMENSION

Passenger satisfaction
for some services
related to the flight

Satisfaction			
id	BInt	PK	
seat_comfort	SInt	NN	
time_convinience	SInt	NN	
food	SInt	NN	
gate_location	SInt	NN	
wifi_service	SInt	NN	
entertainment	SInt	NN	
booking	SInt	NN	
online_support	SInt	NN	
onboard_service	SInt	NN	
leg_room_service	SInt	NN	
bagage_handling	SInt	NN	
online_boarding	SInt	NN	
checkin	SInt	NN	
cleanliness	SInt	NN	

FLIGHT DIMENSION

Flight data

Flight			
id_flight	BInt	PK	
distance	BInt	NN	
departure_delay	BInt		
arrival_delay	BInt		

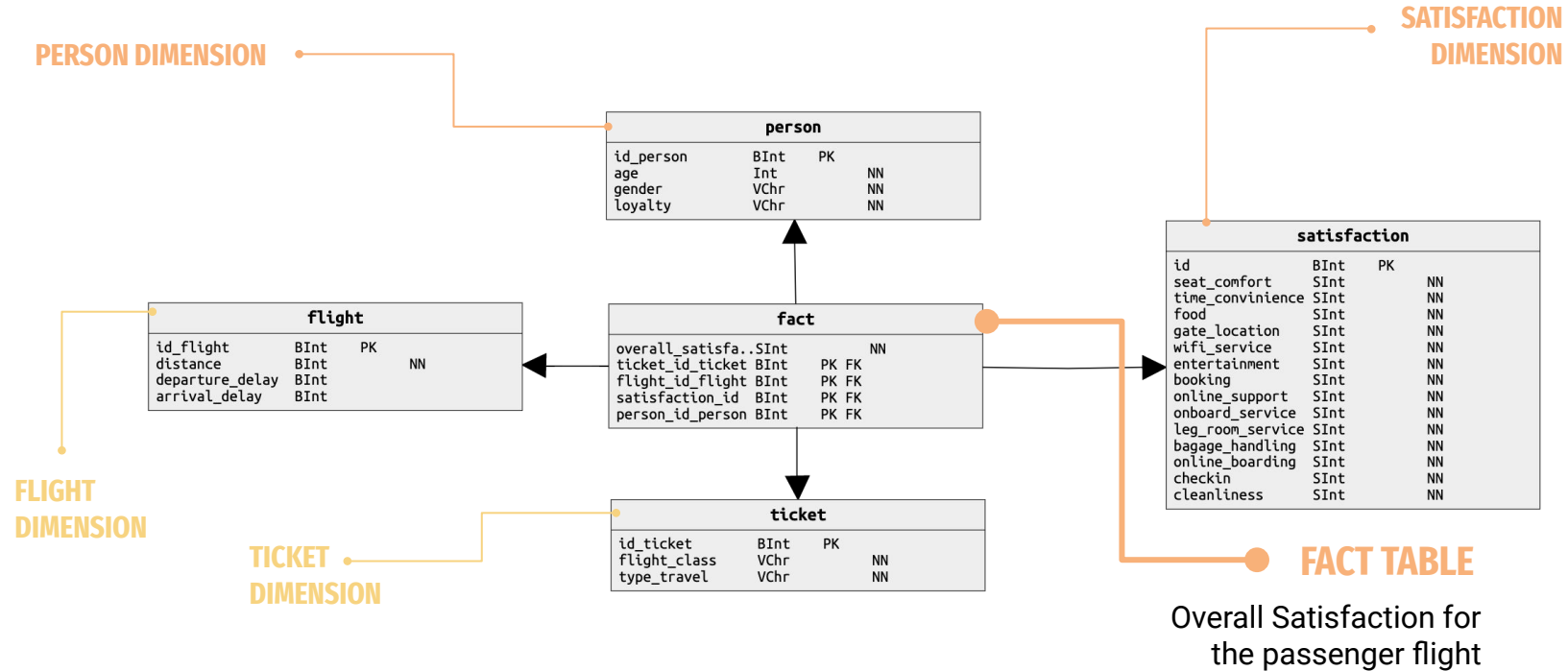
Fact			
overall_satisfa..	SInt	NN	

Ticket			
id_ticket	BInt	PK	
flight_class	VChr	NN	
type_travel	VChr	NN	

TICKET DIMENSION

Informations about the
passenger journey

Data Design



Data Integration



AZURE

Database on Microsoft
data centers



32GB

Storage Capacity

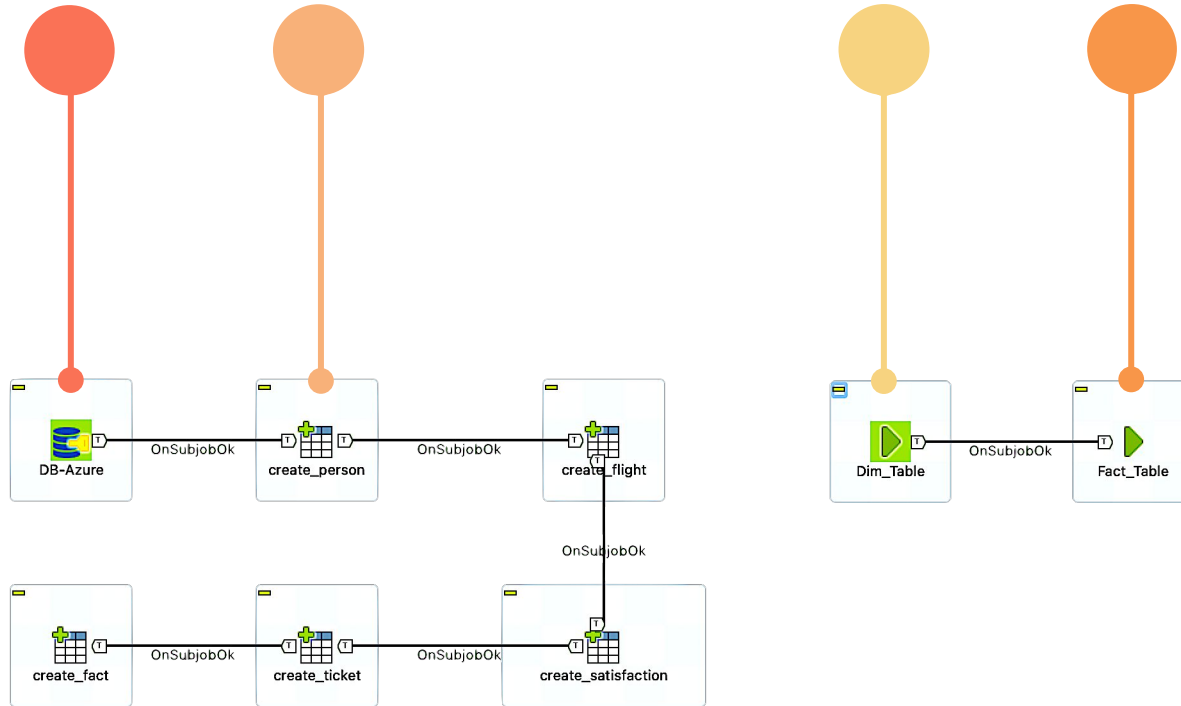
3 zones

Availability

7 days

Backup Retention Period

Data Integration



DB-AZURE

Connection to Azure database

CREATE_PERSON

Creates *Person* table on the database or drop it if exists

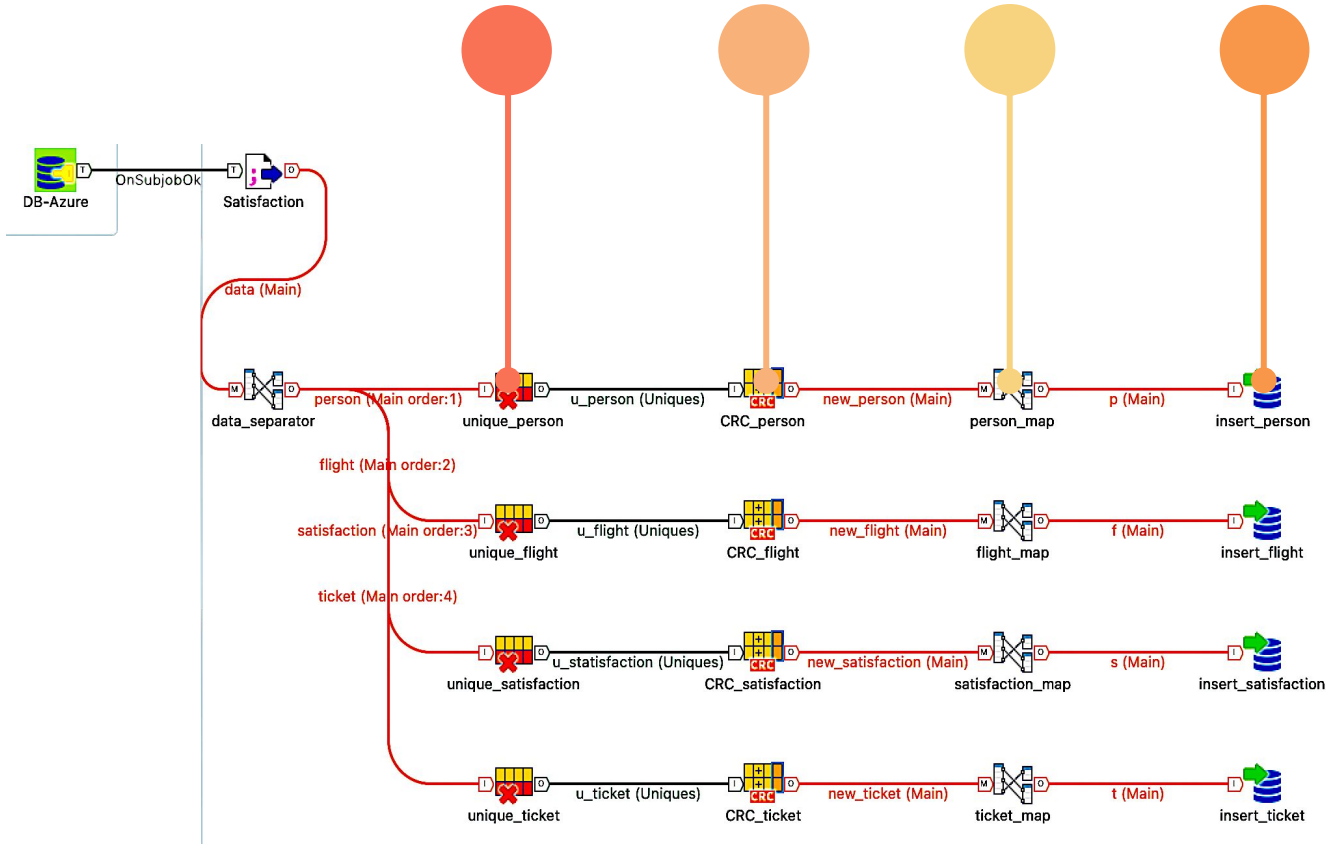
DIM_TABLE

Runs *Dim_Table* job for updating the dimensional tables on the database

FACT_TABLE

Runs *Fact_Table* job for updating the fact table on the database

Data Integration



UNIQUE_PERSON

Outputs unique instances of
person

CRC_PERSON

Generates a unique id for the instances

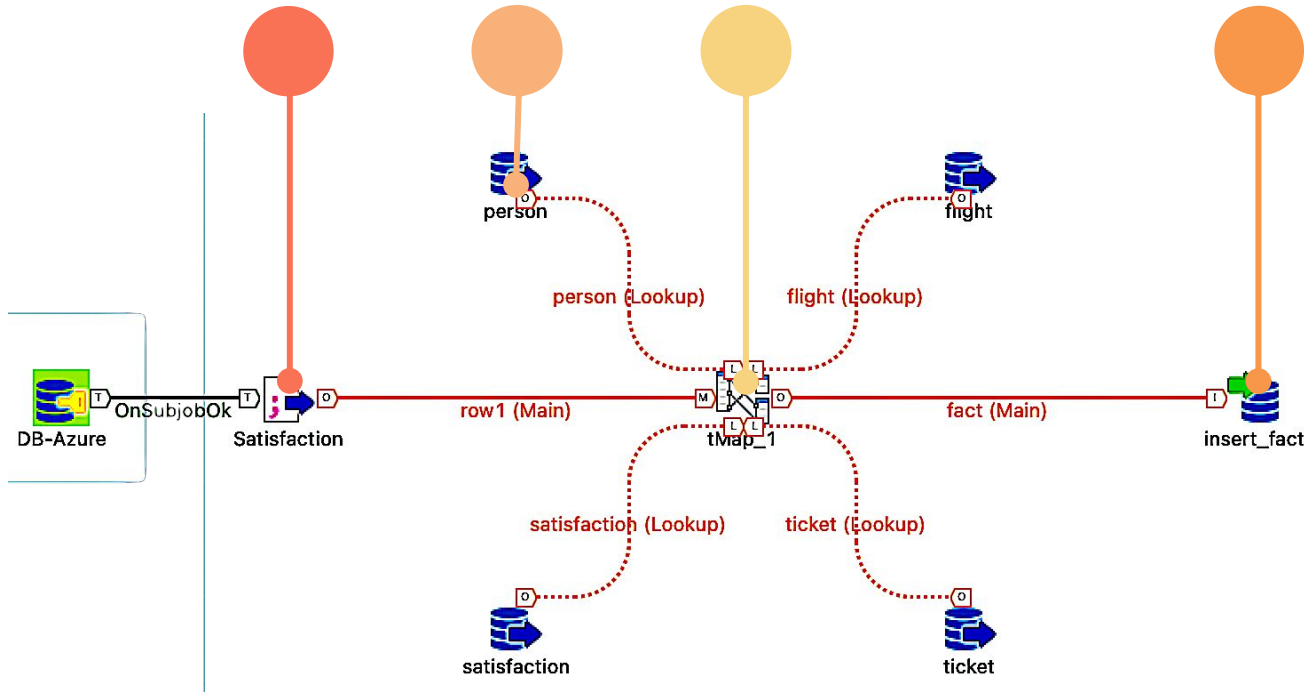
PERSON_MAP

Selects attributes and set their names

INSERT_PERSON

Insert the *person* instance
on the Azure database

Data Integration



SATISFACTION

CSV file with the data

PERSON

Lookup on the table *person* to get all instances

tMAP_1

Joins the data from the lookups in order to find the corresponding instances of the CSV file

INSERT_FACT

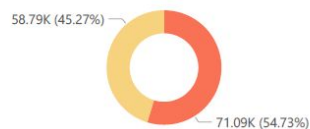
Updates the Azure database with the new records of facts

Data Visualization



Satisfaction

● satisfied ● neutral or dissatisfied



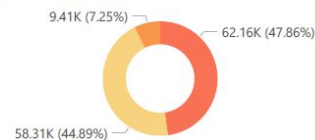
Gender

● Female ● Male



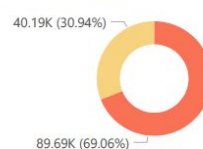
Class

● Business ● Eco ● Eco Plus



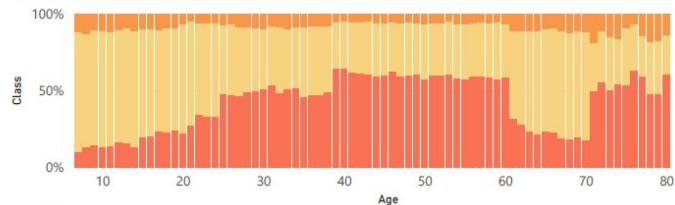
Type of Travel

● Business travel ● Personal Travel



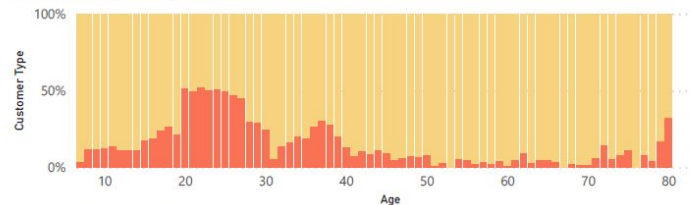
% of Class by Age

Class ● Business ● Eco ● Eco Plus

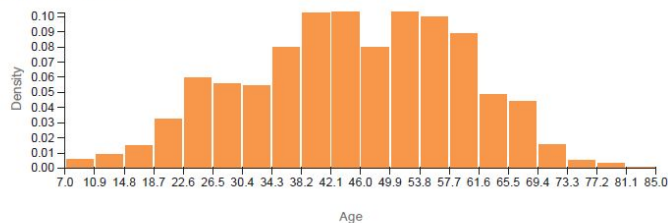


% of Customer Type by Age

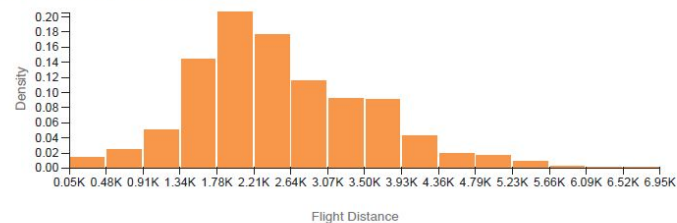
Customer Type ● disloyal Customer ● Loyal Customer



Age Distribution



Flight Distance Distribution



Data Visualization

Average of Checkin service



Average of Inflight entertainment



Average of Cleanliness



Average of Inflight wifi service



Average of Seat comfort



Average of Ease of Online booking



Average of Online support



Average of Gate location



Average of Online boarding



Average of Food and drink



Average of Leg room service



Average of Baggage handling

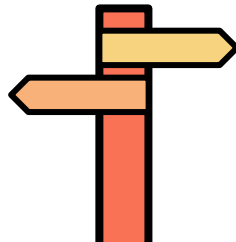


Average of On-board service



What's the average classification
for the provided services?

Which kind of passenger
travels the most?



Who are the most
satisfied customers?

What's the Age and Distance
travelled distribution?

Data Visualization

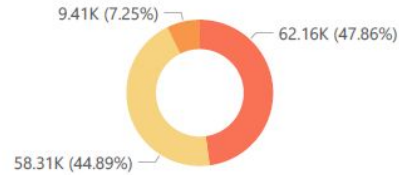
Gender

● Female ● Male



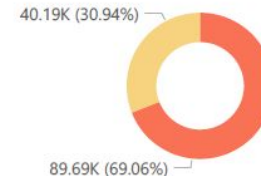
Class

● Business ● Eco ● Eco Plus

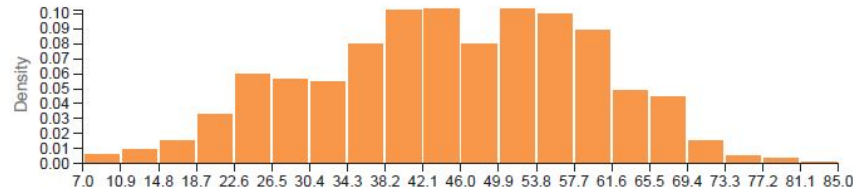


Type of Travel

● Business travel ● Personal Travel

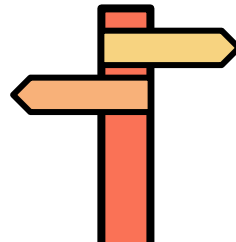


Age Distribution



What's the average classification
for the provided services?

Which kind of passenger
travels the most?



Who are the most
satisfied customers?

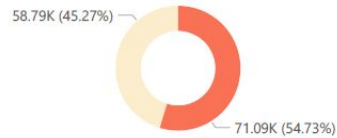
What's the Age and Distance
travelled distribution?

Data Visualization



Satisfaction

● satisfied ● neutral or dissatisfied



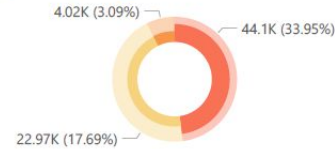
Gender

● Female ● Male



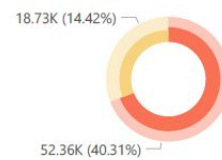
Class

● Business ● Eco ● Eco Plus



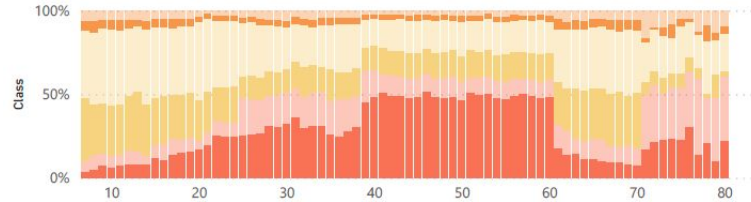
Type of Travel

● Business travel ● Personal Travel



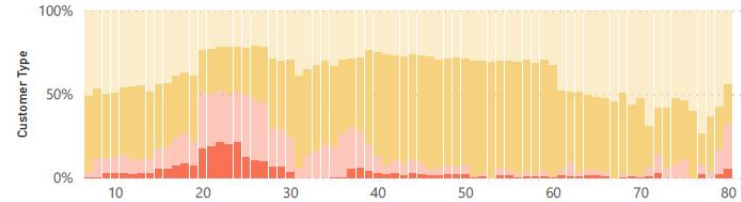
% of Class by Age

Class ● Business ● Eco ● Eco Plus



% of Customer Type by Age

Customer Type ● disloyal Customer ● Loyal Customer



What's the average classification
for the provided services?

Which kind of passenger
travels the most?

Who are the most
satisfied customers?

What's the Age and Distance
travelled distribution?

Data Visualization

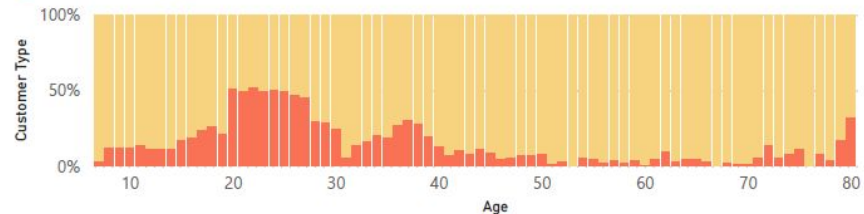
% of Class by Age

Class ● Business ● Eco ● Eco Plus

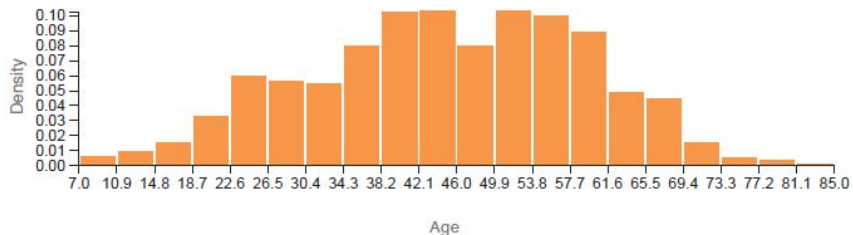


% of Customer Type by Age

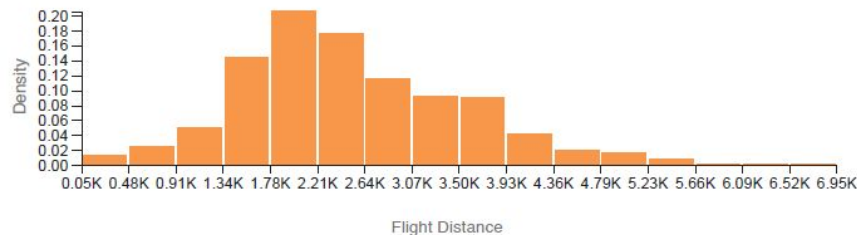
Customer Type ● disloyal Customer ● Loyal Customer



Age Distribution



Flight Distance Distribution



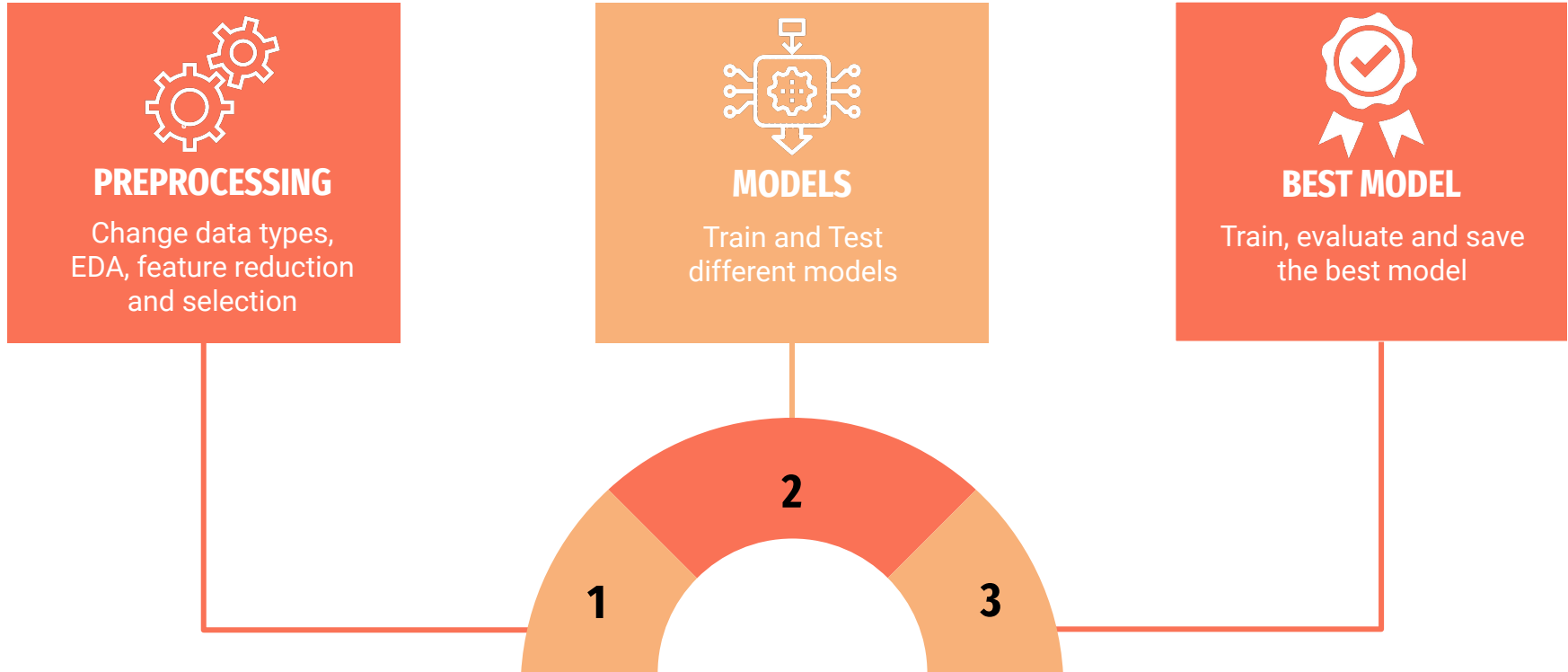
What's the average classification
for the provided services?

Which kind of passenger
travels the most?

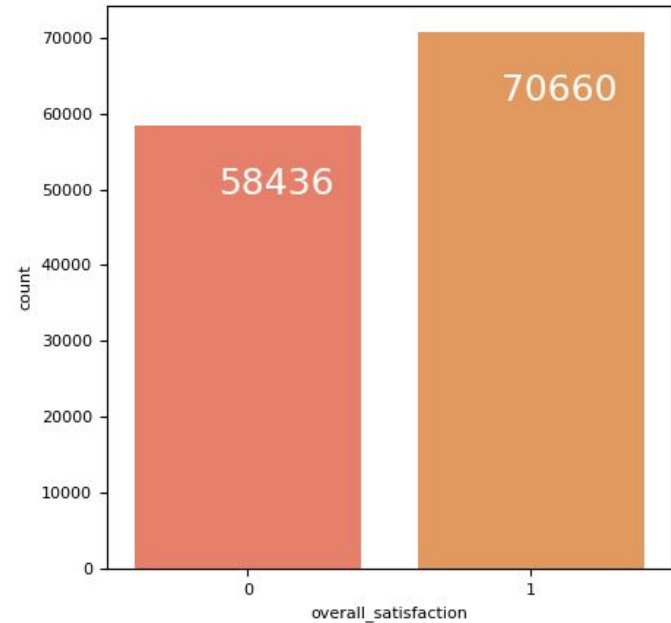
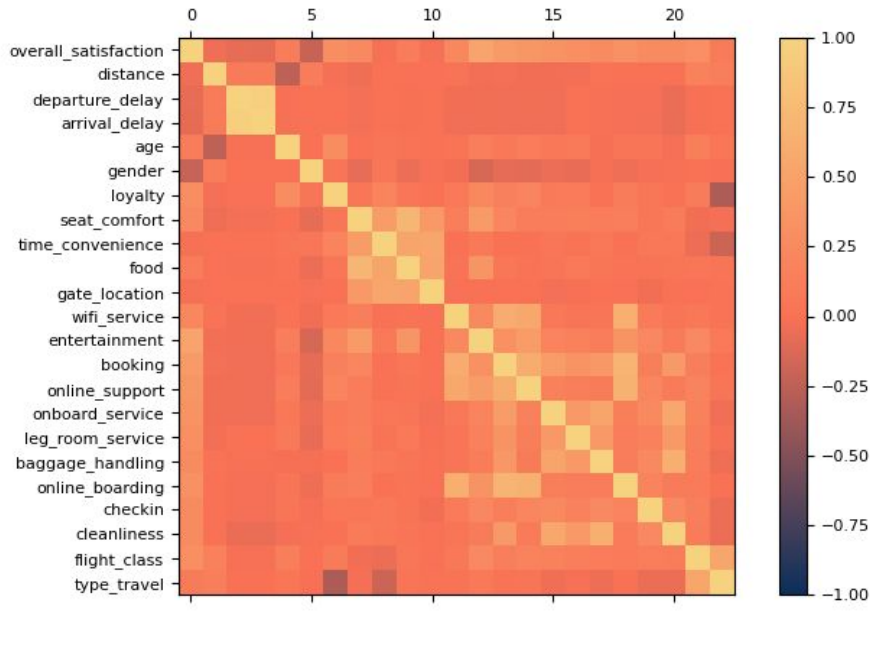
Who are the most
satisfied customers?

What's the Age and Distance
travelled distribution?

Machine Learning



Exploratory Data Analysis



Feature Selection / Reduction



Univariate Selection

Univariate feature selection works by selecting the best features based on univariate statistical tests

Recursive Feature Elimination

Select features by recursively considering smaller and smaller sets of features

Principal Component Analysis

Linear dimensionality reduction using Singular Value Decomposition of the data to project it to a lower dimensional space

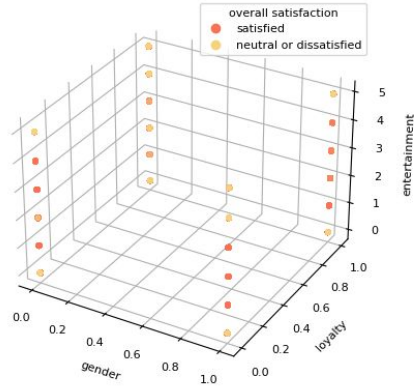
Feature Importance

The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature.

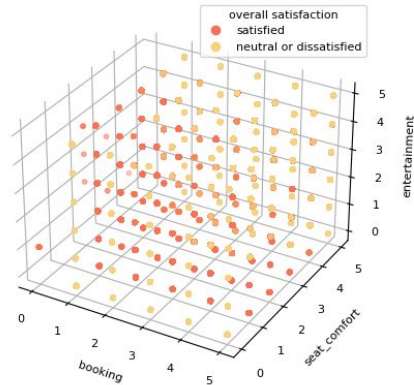
Since we only have **22** features we tried reducing to **10** and **3** features.

Feature Selection / Reduction - 3 Dimensions

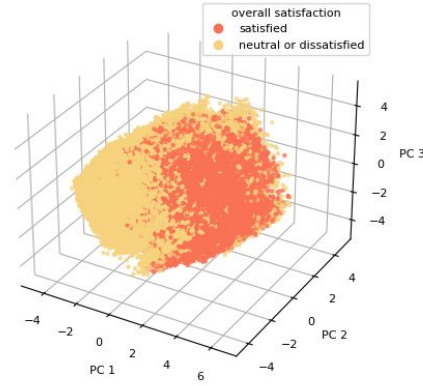
**Recursive
Feature
Elimination**



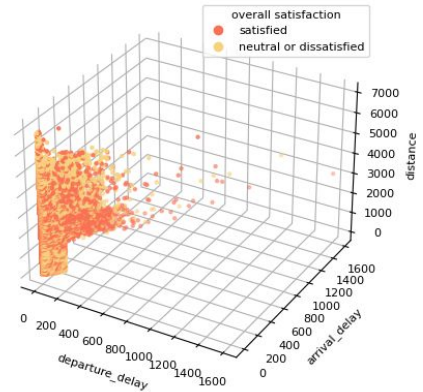
**Feature
Importance**



**Principal
Component
Analysis**

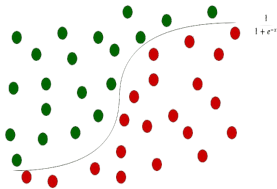


**Univariate
Selection**

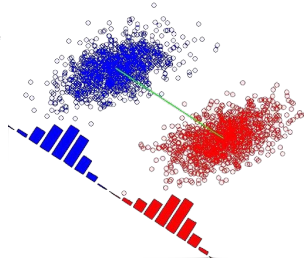


Classification Models

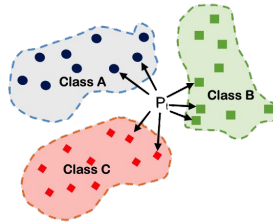
Logistic Regression



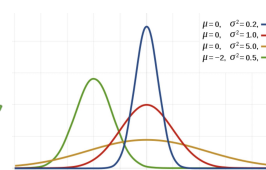
Linear Discriminant Analysis



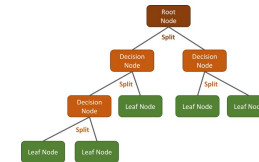
K-Nearest Neighbors



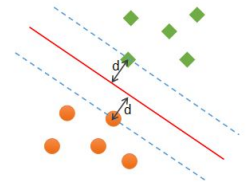
Gaussian Naive Bayes



Decision Tree Classifier (CART)



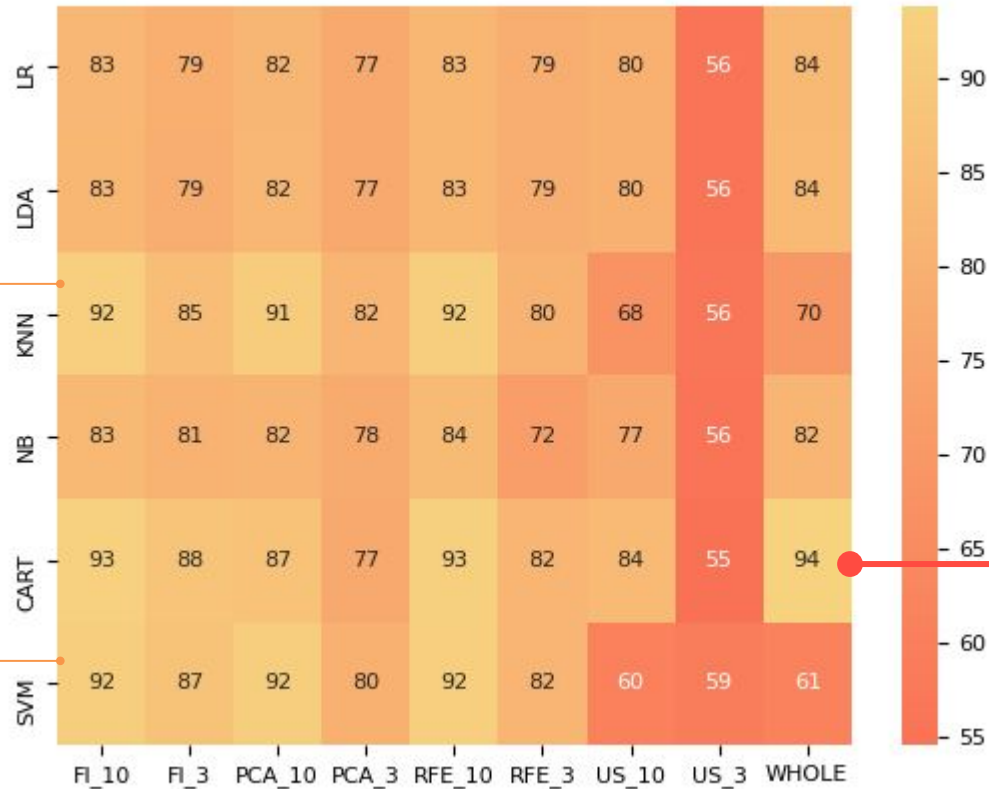
Support Vector Machine





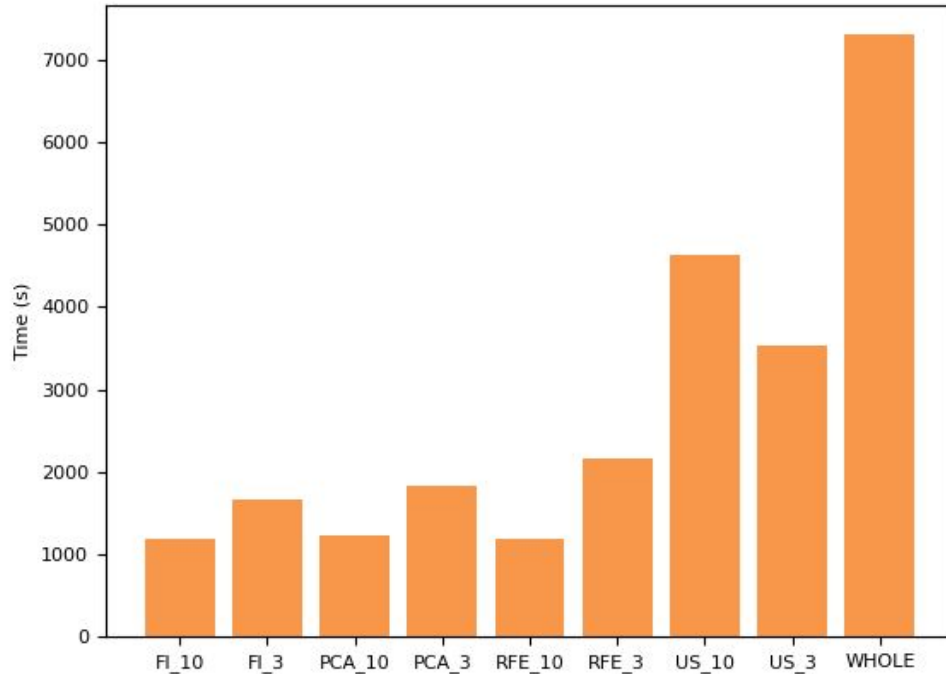
Comparing Accuracy Means

ALSO GOOD RESULTS



BEST MODEL

Comparing Times : SVM Example



WHOLE

All features



US_X

Univariate
Statistical tests
with 3 or 10
features



RFE_X

Recursive
Feature
Elimination with
3 or 10 features



PCA_X

Principal
Component
Analysis with 3
or 10 features



FI_X

Feature
Importance with
3 or 10 features

Best Model : CART - Whole

93.7%



ACCURACY

What proportion of identifications was actually correct?

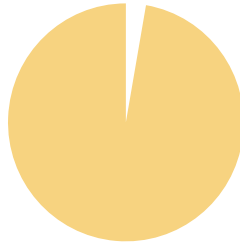
94.0%



PRECISION

What proportion of positive identifications was actually correct?

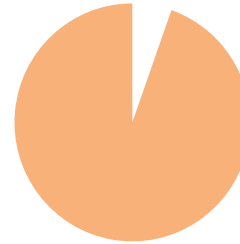
94.4%



SENSITIVITY

What proportion of actual positives was identified correctly?

92.7%



SPECIFICITY

What proportion of actual negatives was identified correctly?

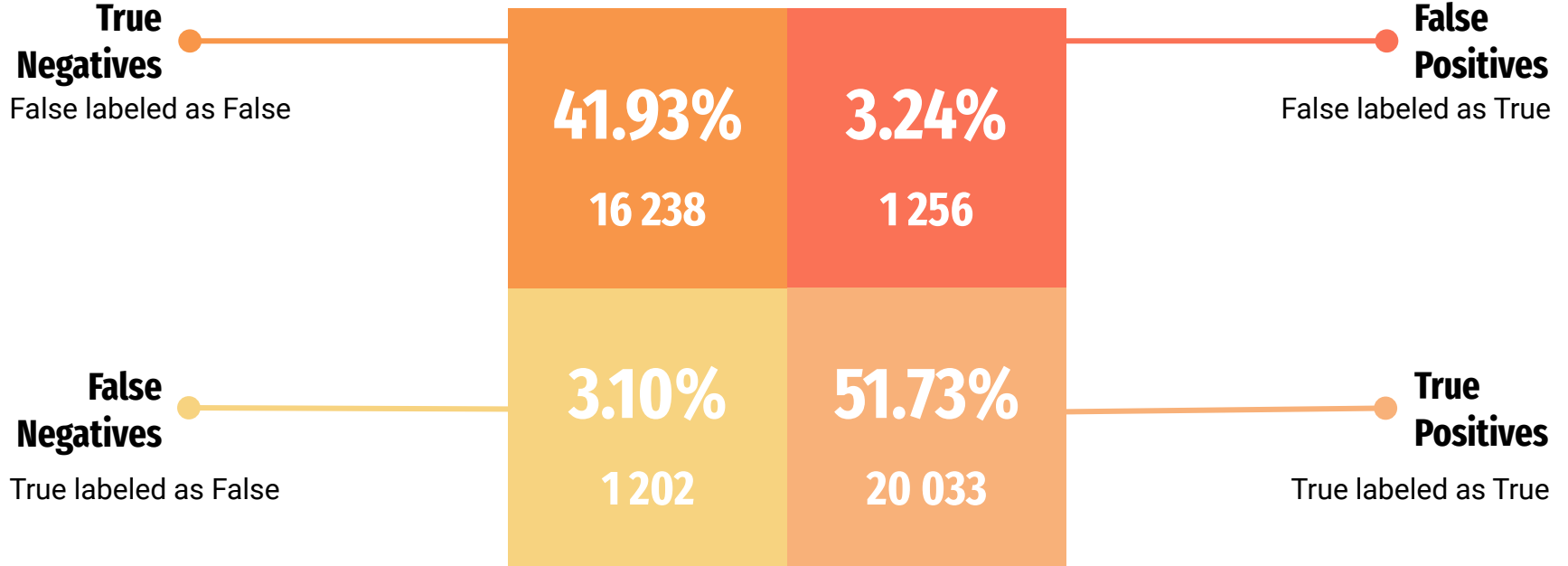
93.5%



AUC SCORE

How much the model is capable of distinguishing between classes

Confusion Matrix



Thanks for your Attention!

Any Questions?

