US Airline Passenger Satisfaction

Business Intelligence, May 2022

Our Team

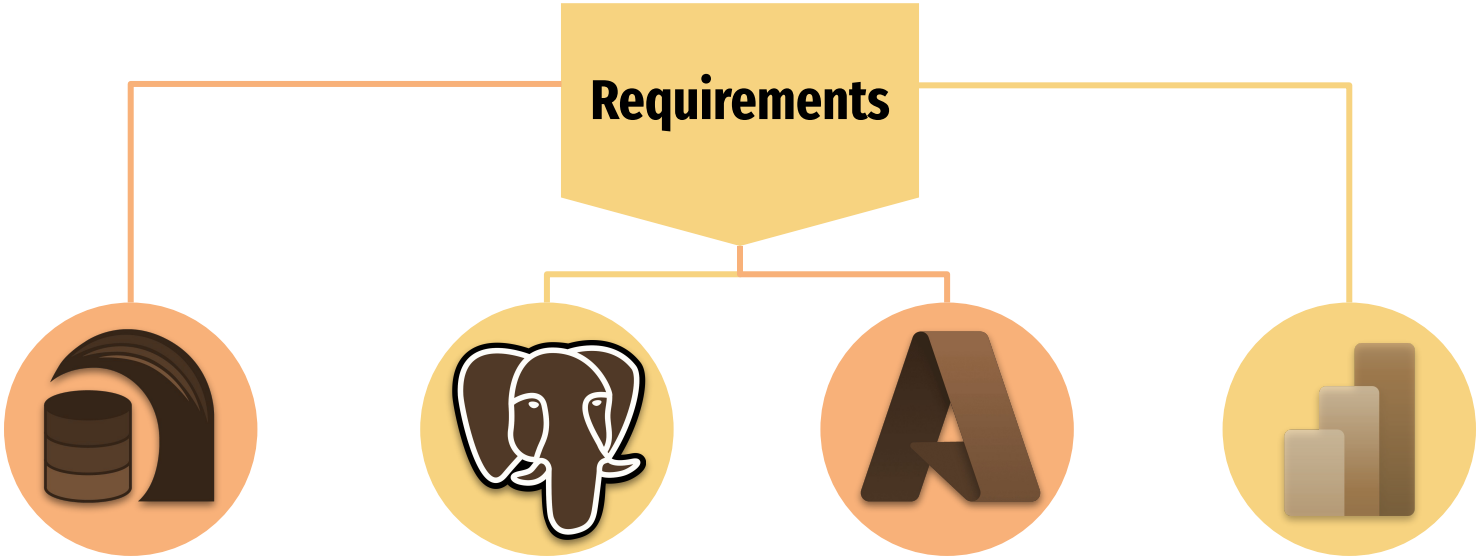
Bruno Faria

2018295474
brunofaria@student.dei.uc.pt

Dylan Perdigão

2018233092
dgp@student.dei.uc.pt

Requirements



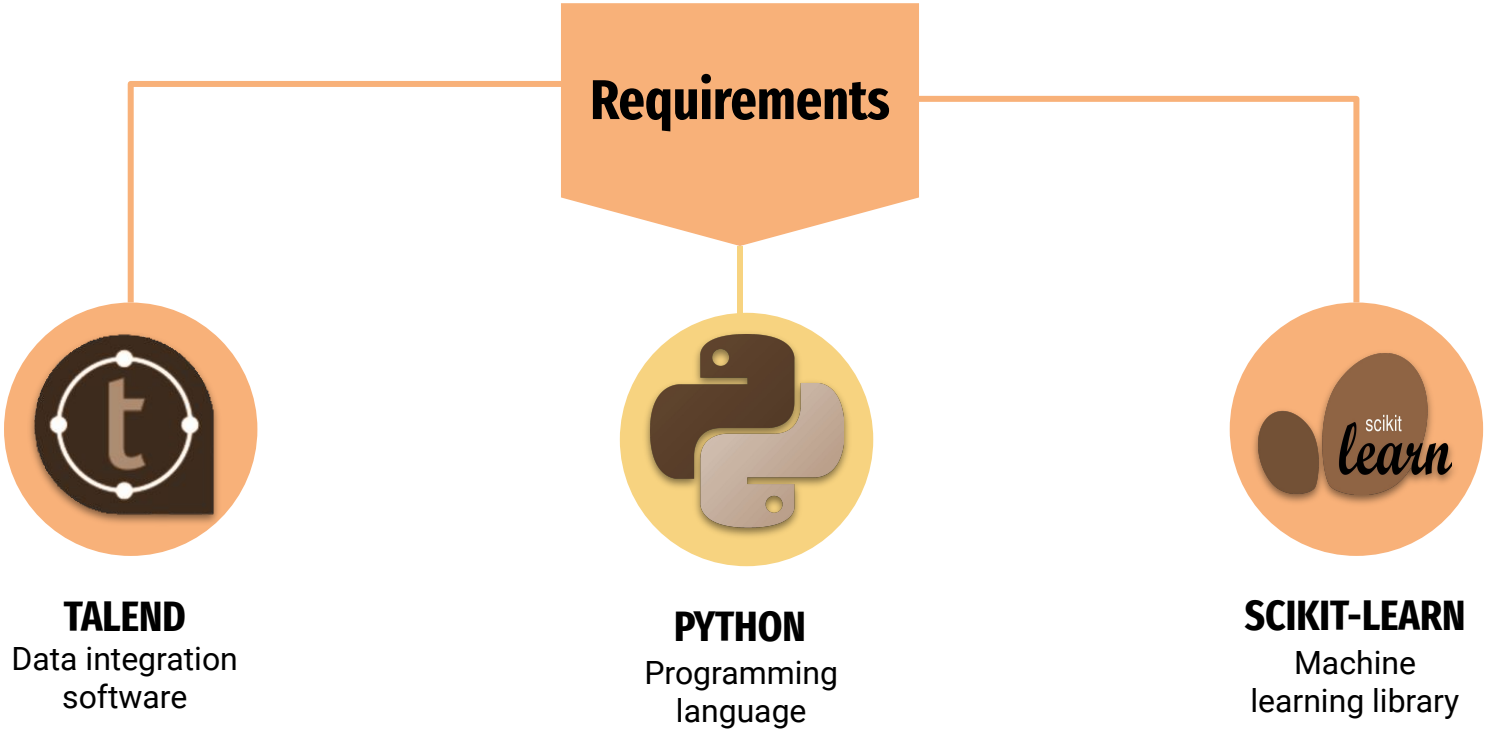
ONDA
Database Modeling

POSTGRESQL
Relational database management system

AZURE
Cloud computing service

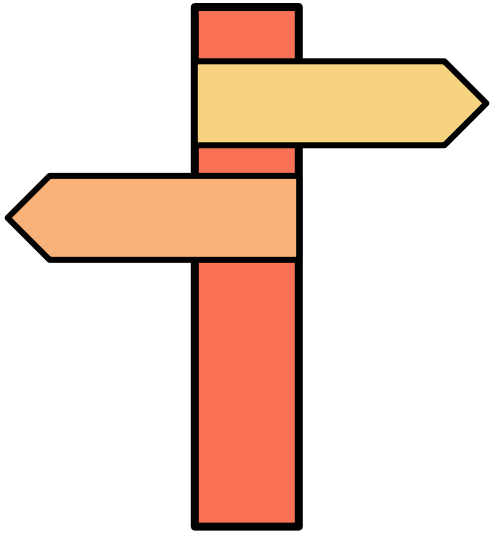
POWER BI
Interactive data visualization software

Requirements



Goals

What's the average classification for the provided services?



Who are the most satisfied customers?

Which kind of passenger travels the most?

What's the Age and Distance travelled distribution?

Predict the satisfaction using machine learning techniques

Dataset

Provenance

Kaggle (see here)

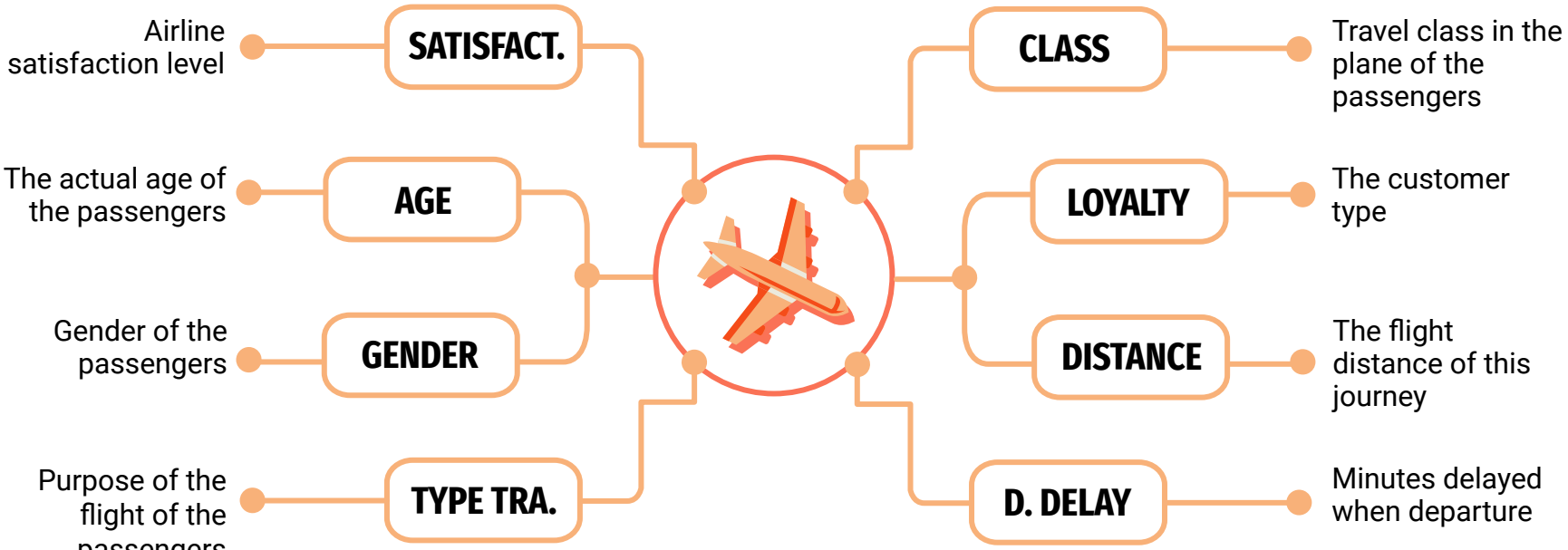
Rows

129880 instances

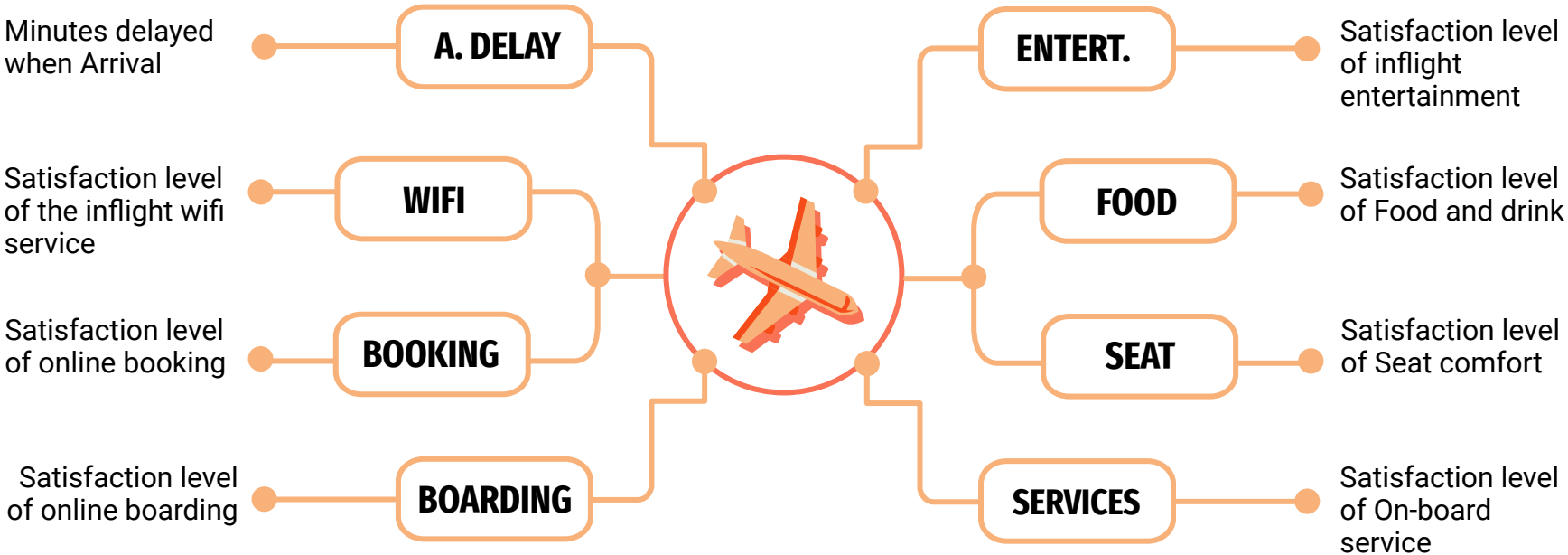
Columns

23 columns

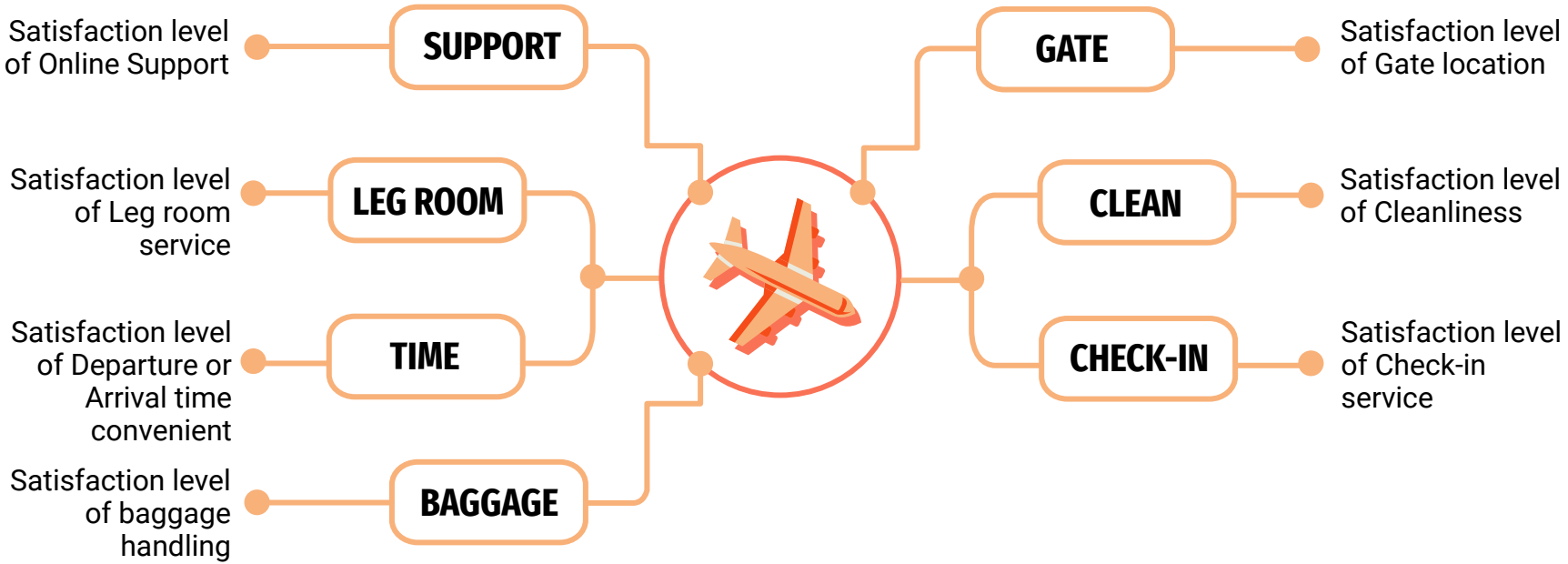
Dataset


Airline satisfaction level
SATISFACT.
The actual age of the passengers
AGE
Gender of the passengers
GENDER
Purpose of the flight of the passengers
TYPE TRA.
CLASS
Travel class in the plane of the passengers
LOYALTY
The customer type
DISTANCE
The flight distance of this journey
D. DELAY
Minutes delayed when departure

Dataset


Minutes delayed when Arrival
A. DELAY
Satisfaction level of the inflight wifi service
WIFI
Satisfaction level of online booking
BOOKING
Satisfaction level of online boarding
BOARDING
ENTERT.
Satisfaction level of inflight entertainment
FOOD
Satisfaction level of Food and drink
SEAT
Satisfaction level of Seat comfort
SERVICES
Satisfaction level of On-board service

Dataset


Satisfaction level of Online Support
SUPPORT
Satisfaction level of Leg room service
LEG ROOM
Satisfaction level of Departure or Arrival time convenient
TIME
Satisfaction level of baggage handling
BAGGAGE
GATE
Satisfaction level of Gate location
CLEAN
Satisfaction level of Cleanliness
CHECK-IN
Satisfaction level of Check-in service

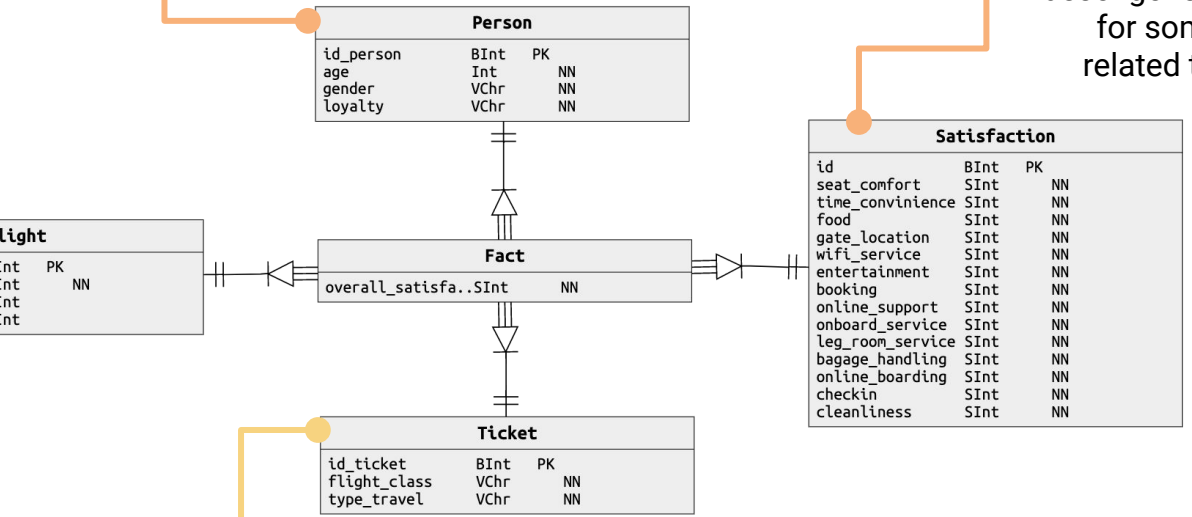
Data Design

PERSON DIMENSION

Passenger basic personal data

SATISFACTION DIMENSION

Passenger satisfaction for some services related to the flight



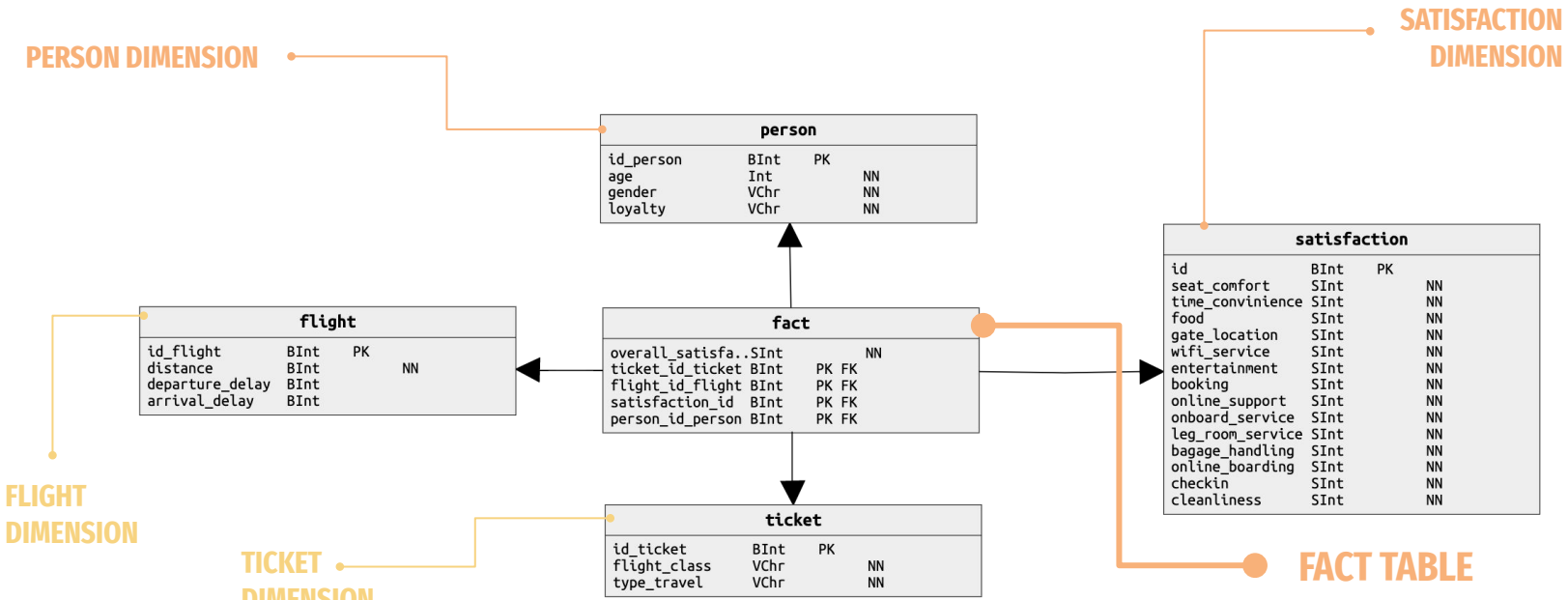
FLIGHT DIMENSION

Flight data

TICKET DIMENSION

Informations about the passenger journey

Data Design


PERSON DIMENSION
SATISFACTION DIMENSION
person
id_person BInt PK
age Int NN
gender VChr NN
loyalty VChr NN
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
id_flight BInt PK
distance BInt NN
departure_delay BInt
arrival_delay BInt
fact
overall_satisfa..SInt NN
ticket_id_ticket BInt PK FK
flight_id_flight BInt PK FK
satisfaction_id BInt PK FK
person_id_person BInt PK FK
FLIGHT DIMENSION
TICKET DIMENSION
ticket
id_ticket BInt PK
flight_class VChr NN
type_travel VChr NN
FACT TABLE
Overall Satisfaction for the passenger flight

Data Integration

AZURE

Database on Microsoft data centers

32GB

Storage Capacity

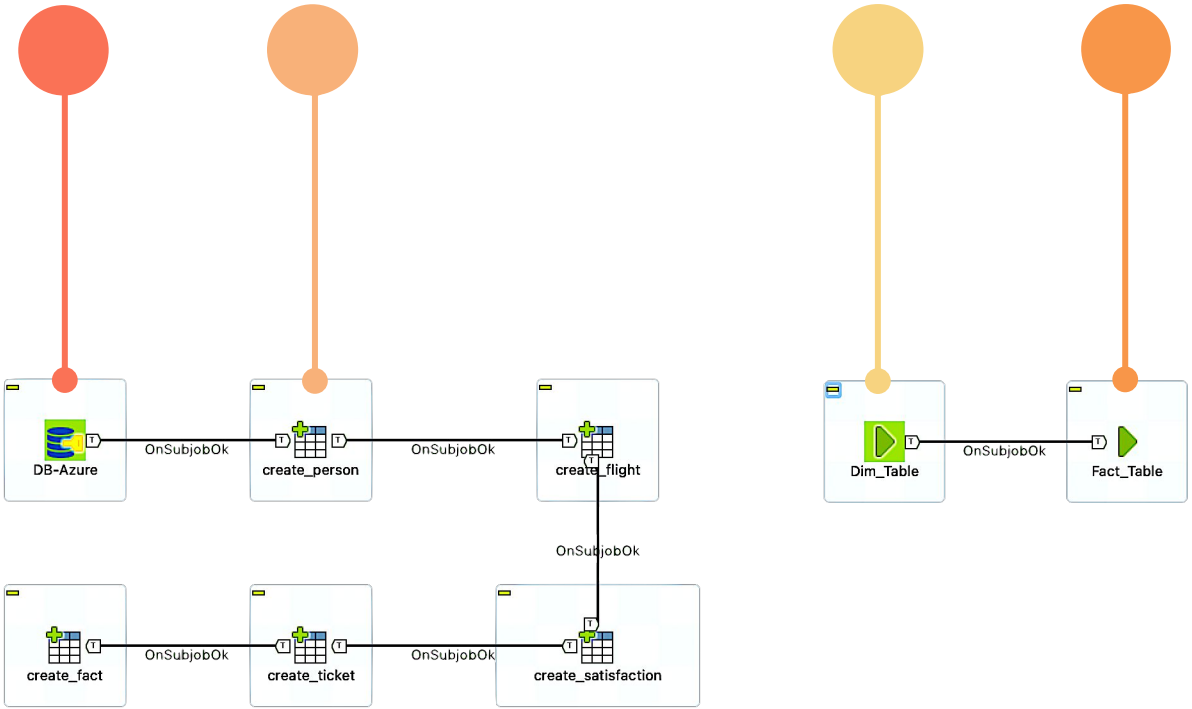
3 zones

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

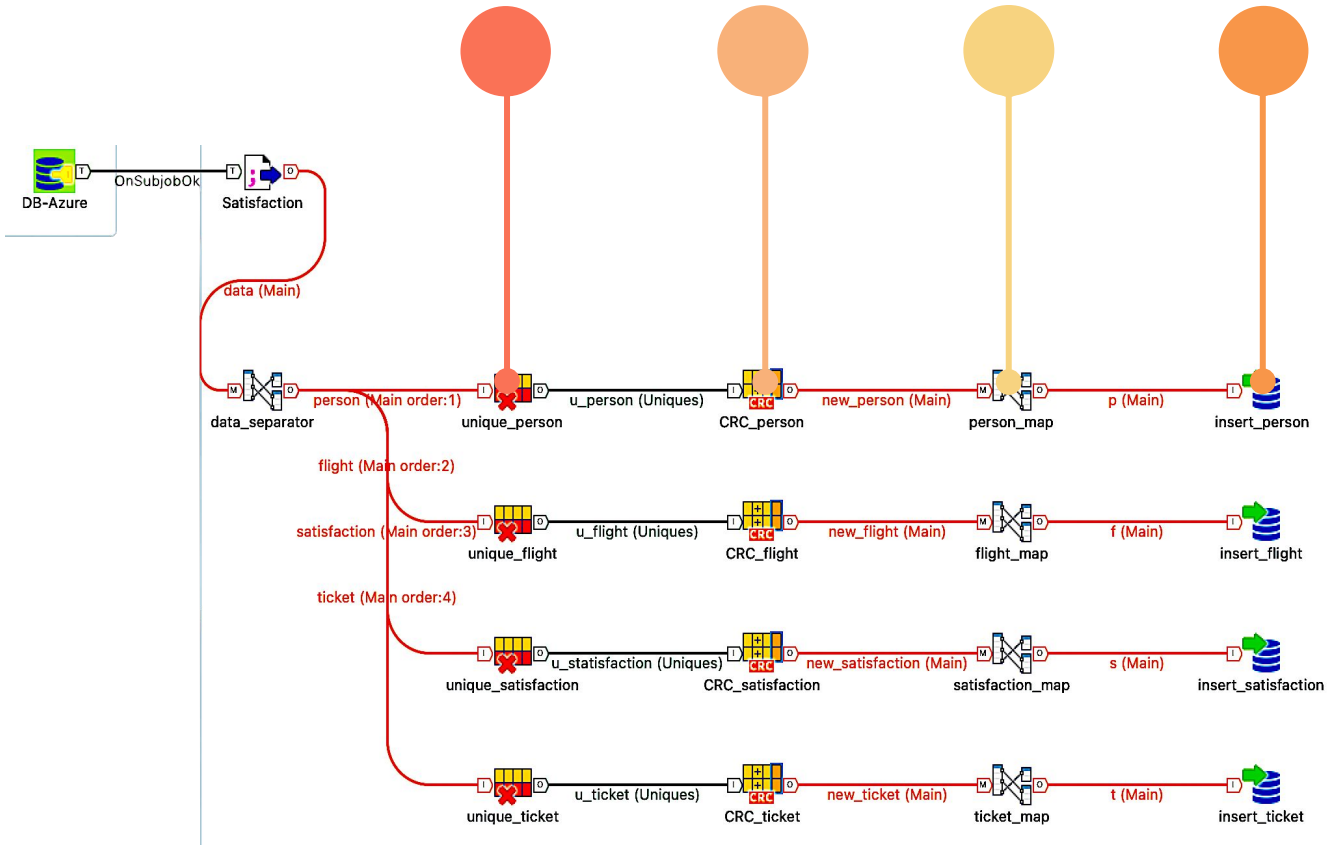
DIM_TABLE

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

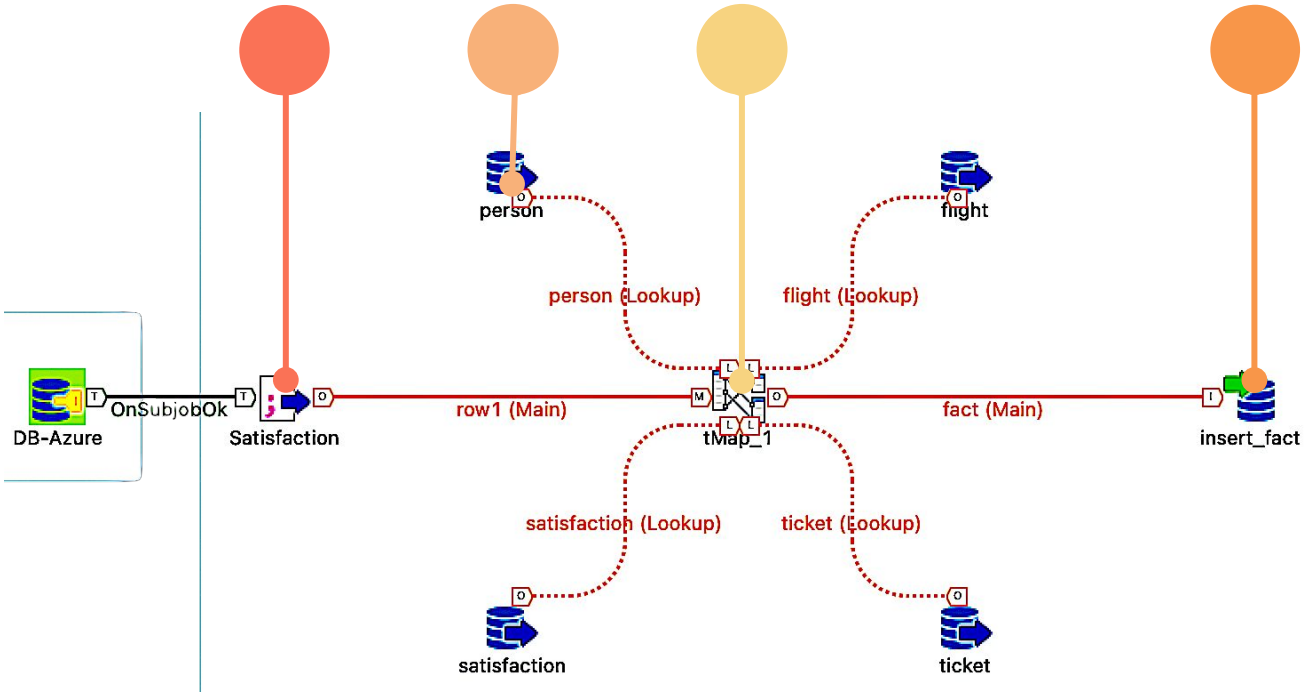
PERSON_MAP

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

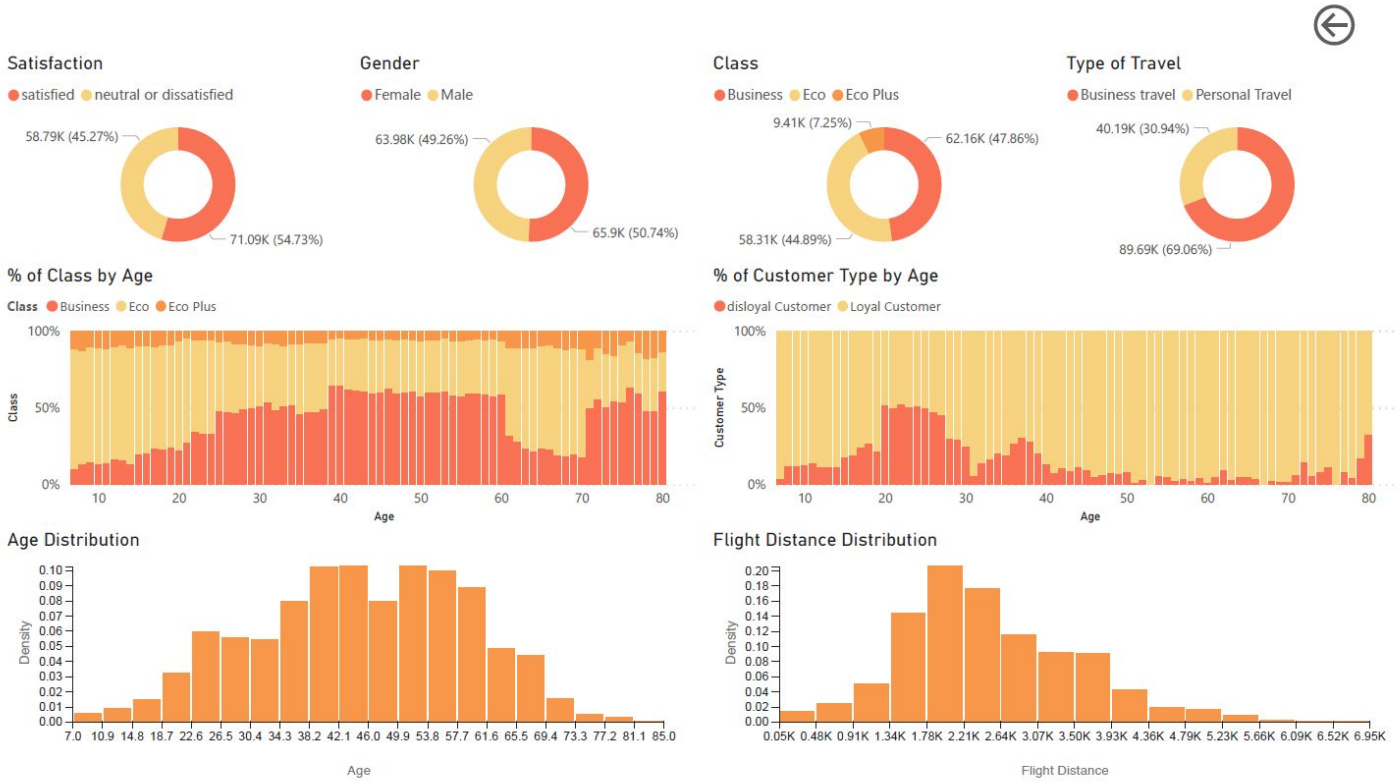
tMAP_1

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

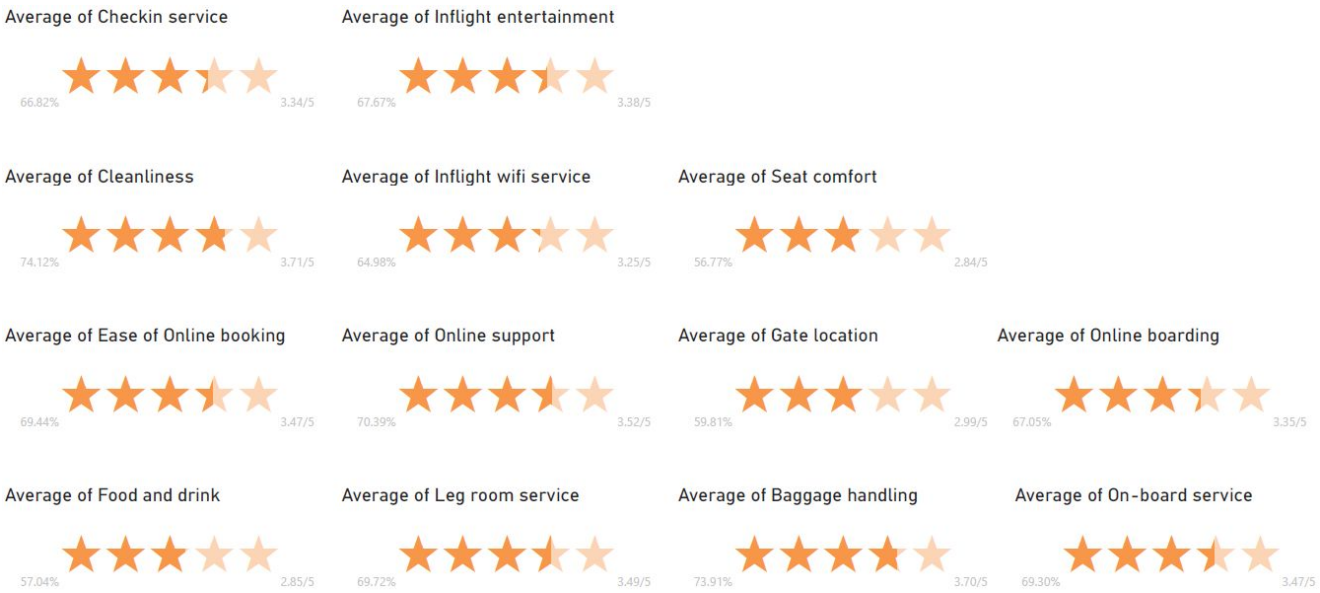
INSERT_FACT

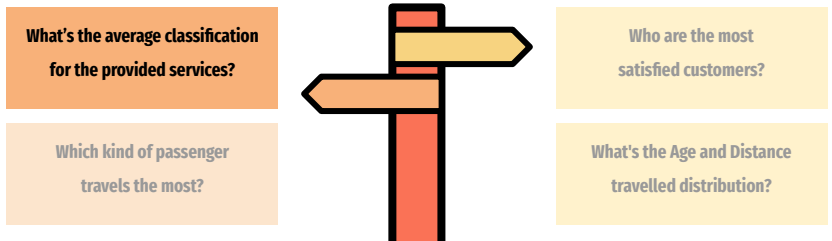
Updates the Azure database with the new records of facts

Data Visualization

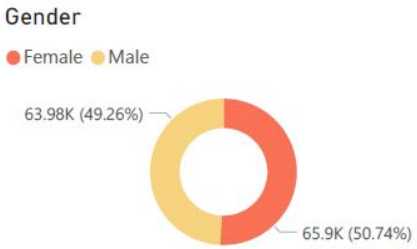


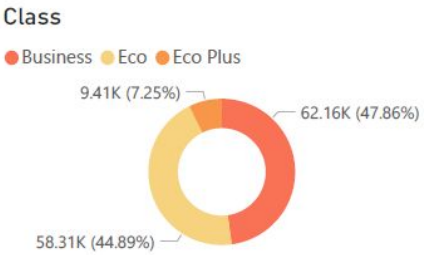
Data Visualization

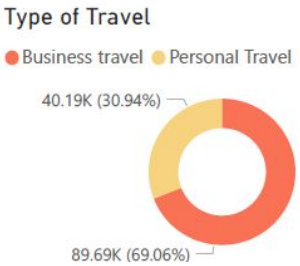


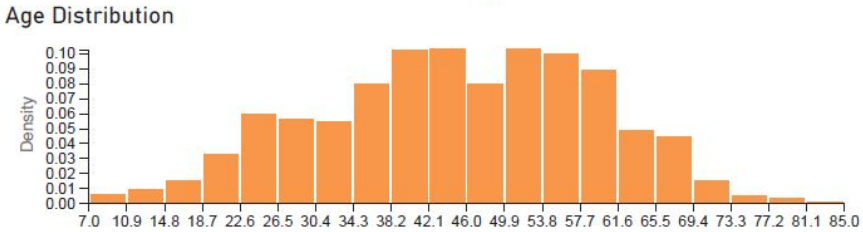


Data Visualization







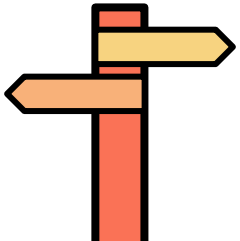


What's the average classification for the provided services?

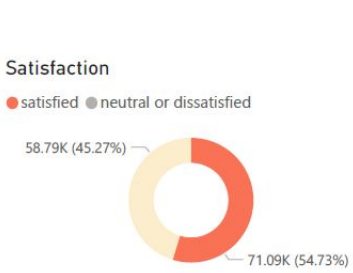
Who are the most satisfied customers?

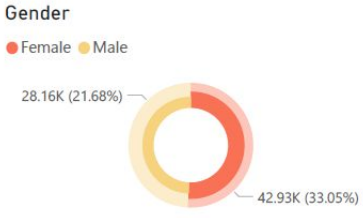
Which kind of passenger travels the most?

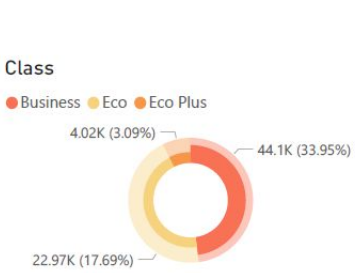
What's the Age and Distance travelled distribution?

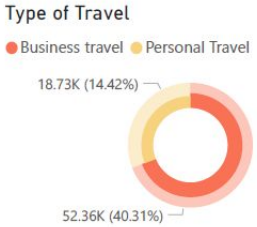


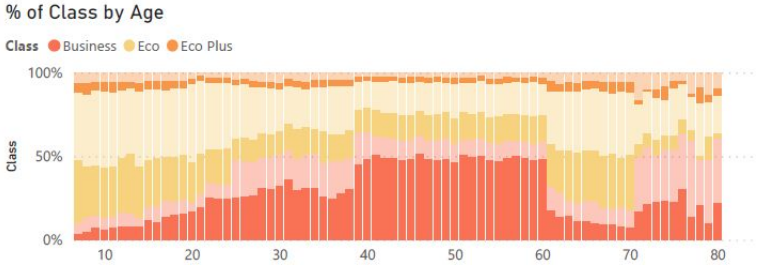
Data Visualization

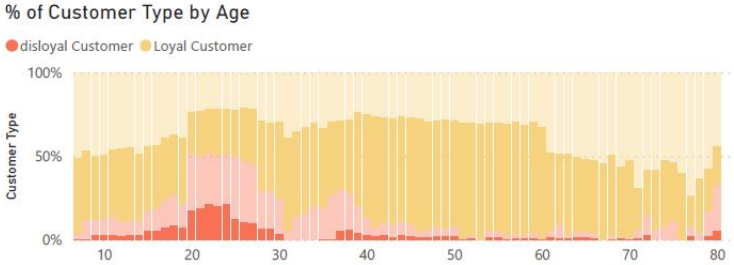






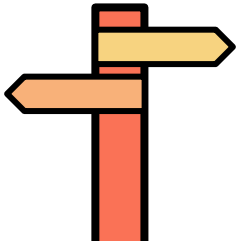






What's the average classification for the provided services?

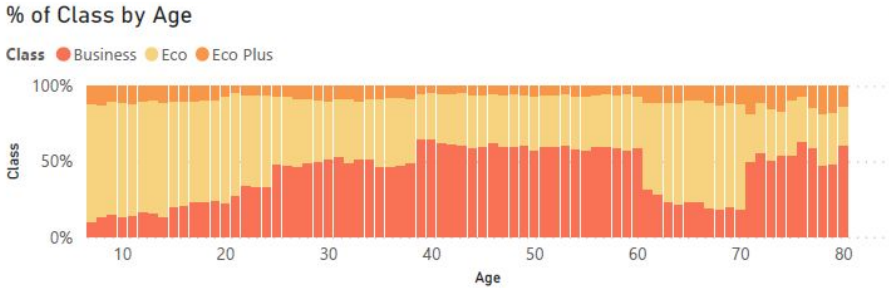
Which kind of passenger travels the most?

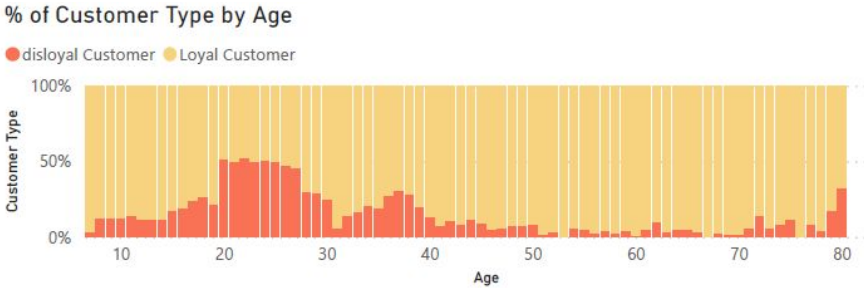


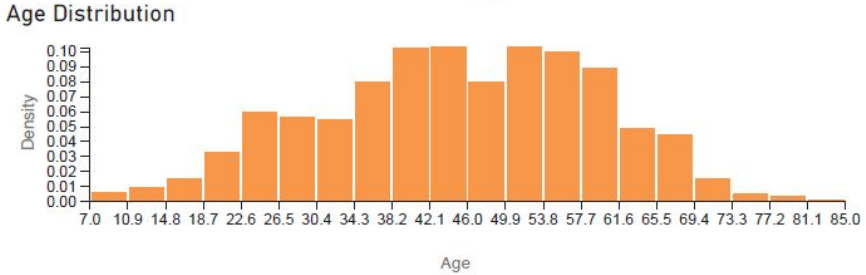
Who are the most satisfied customers?

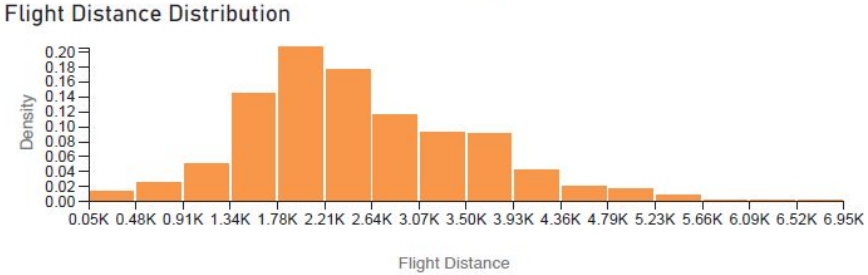
What's the Age and Distance travelled distribution?

Data Visualization









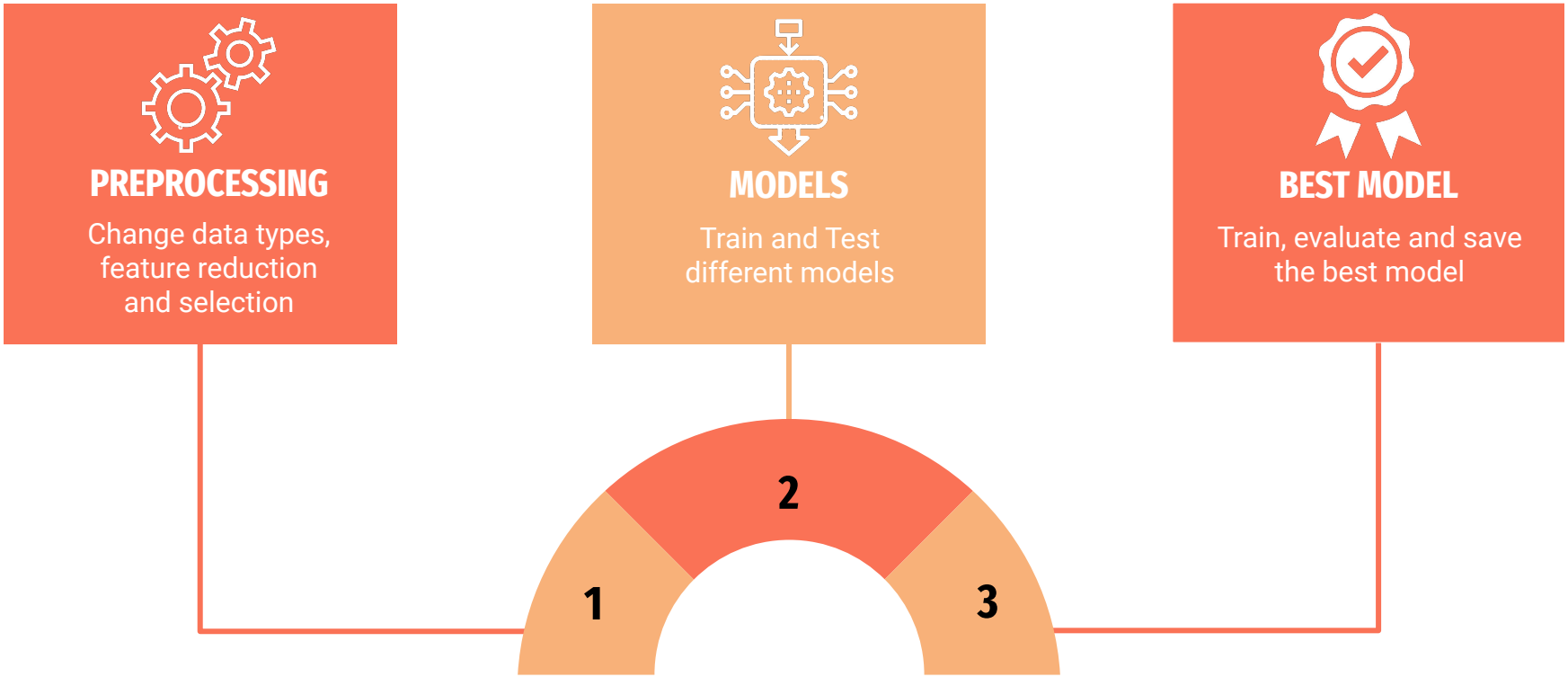
What's the average classification for the provided services?

Who are the most satisfied customers?

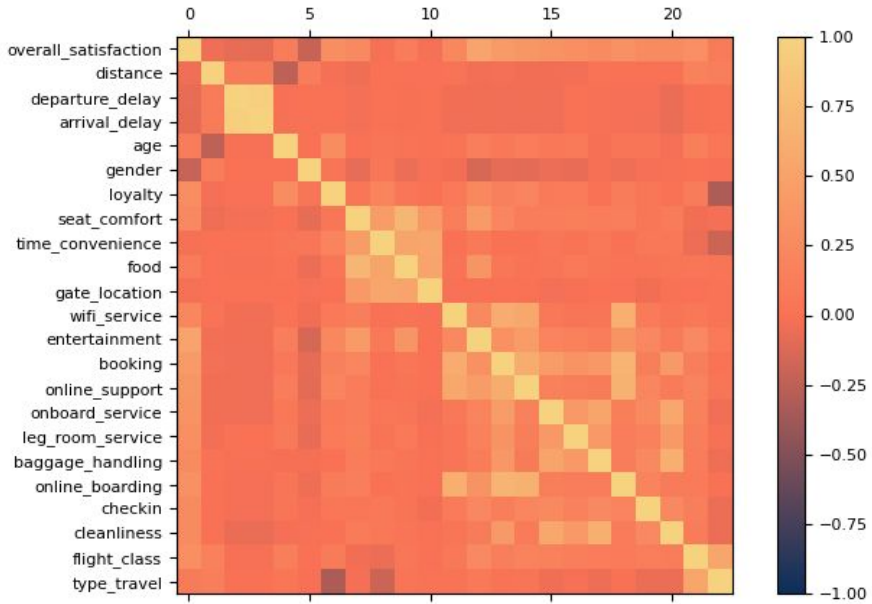
Which kind of passenger travels the most?

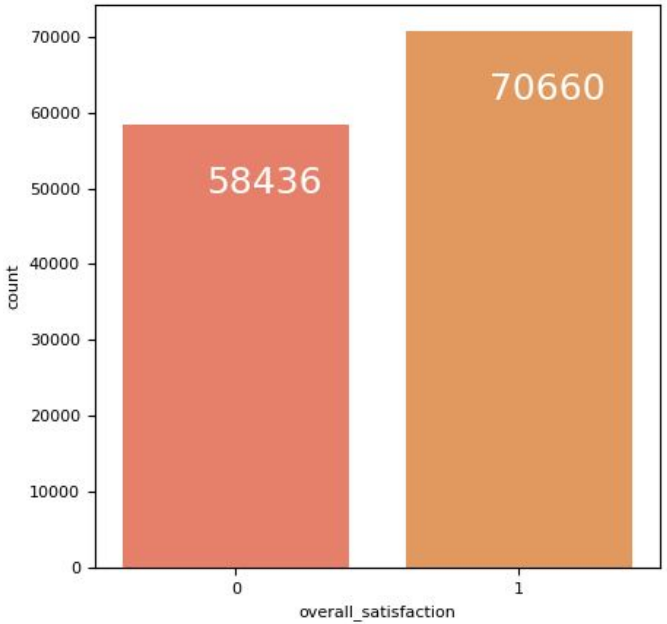
What's the Age and Distance travelled distribution?

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

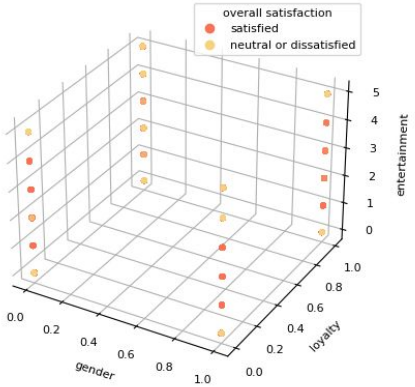
Feature Importance

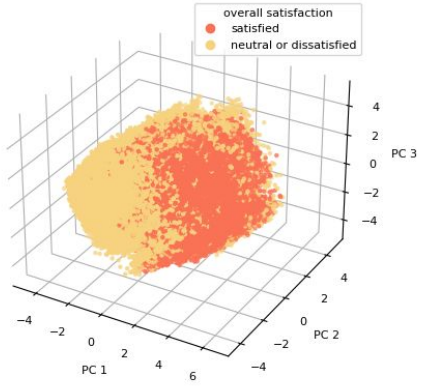
The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature.

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

Feature Selection / Reduction - 3 Dimensions

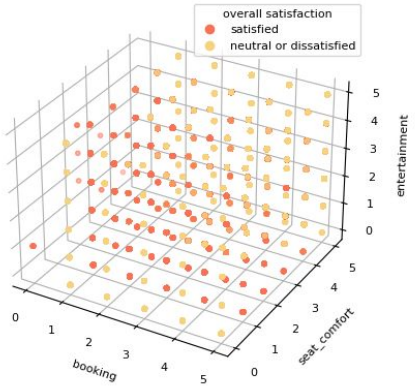
Recursive Feature Elimination

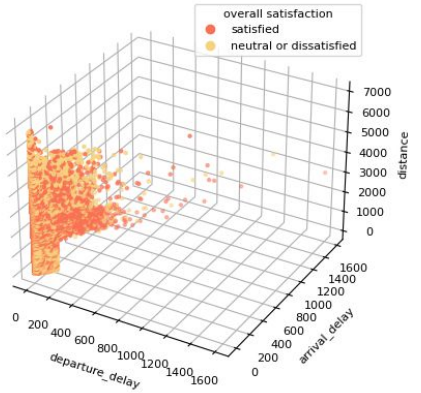




Principal Component Analysis

Feature Importance

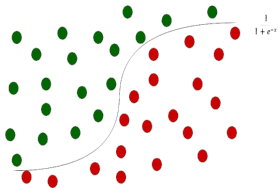


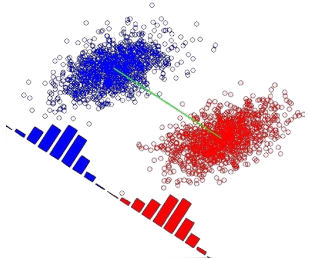


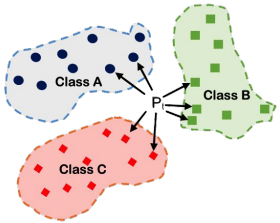
Univariate Selection

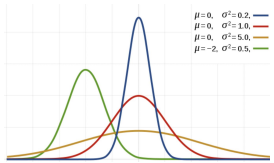
Classification / Regression Models

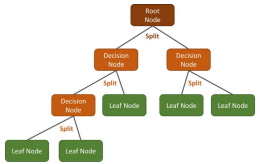
- Logistic Regression
- Linear Discriminant Analysis
- K-Nearest Neighbors
- Gaussian Naive Bayes
- Decision Tree Classifier (CART)
- Support Vector Machine

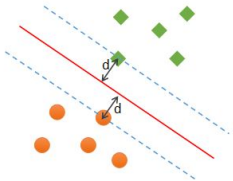




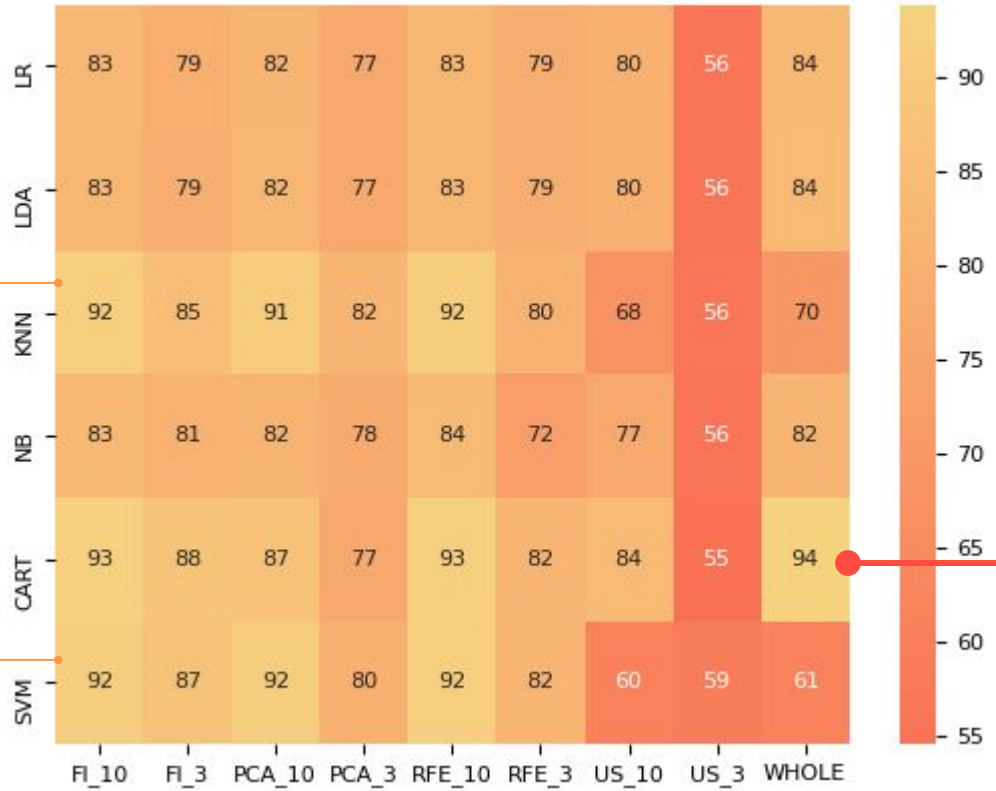








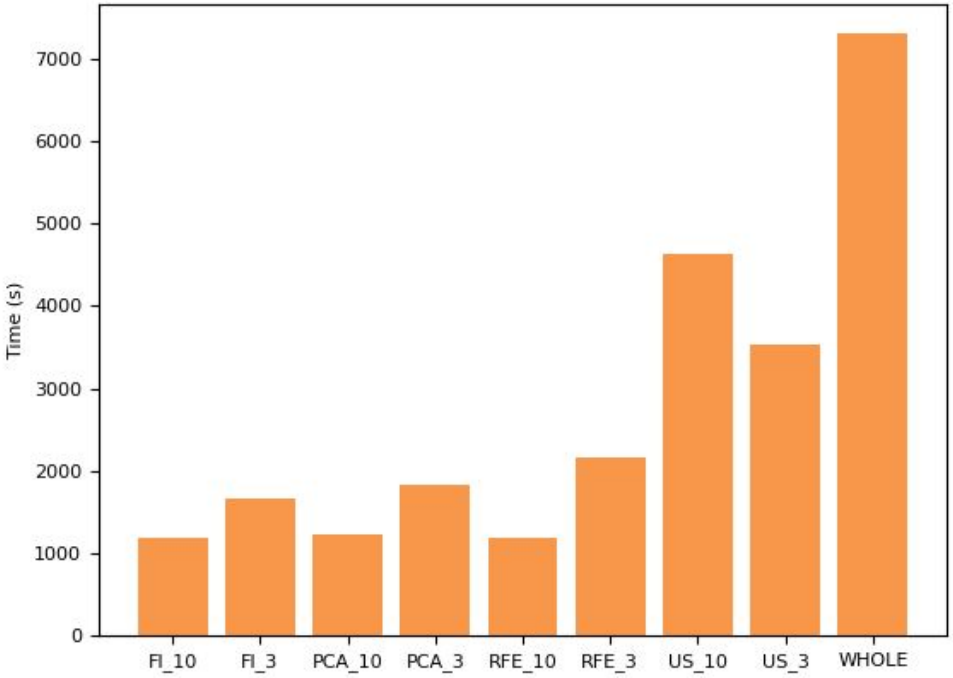
Comparing Accuracy Means



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

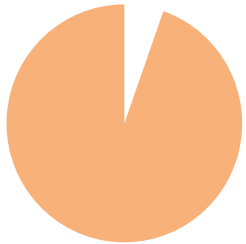
What proportion of positive identifications was actually correct?

94.4%

SENSITIVITY

What proportion of actual positives was identified correctly?

92.7%



SPECIFICITY

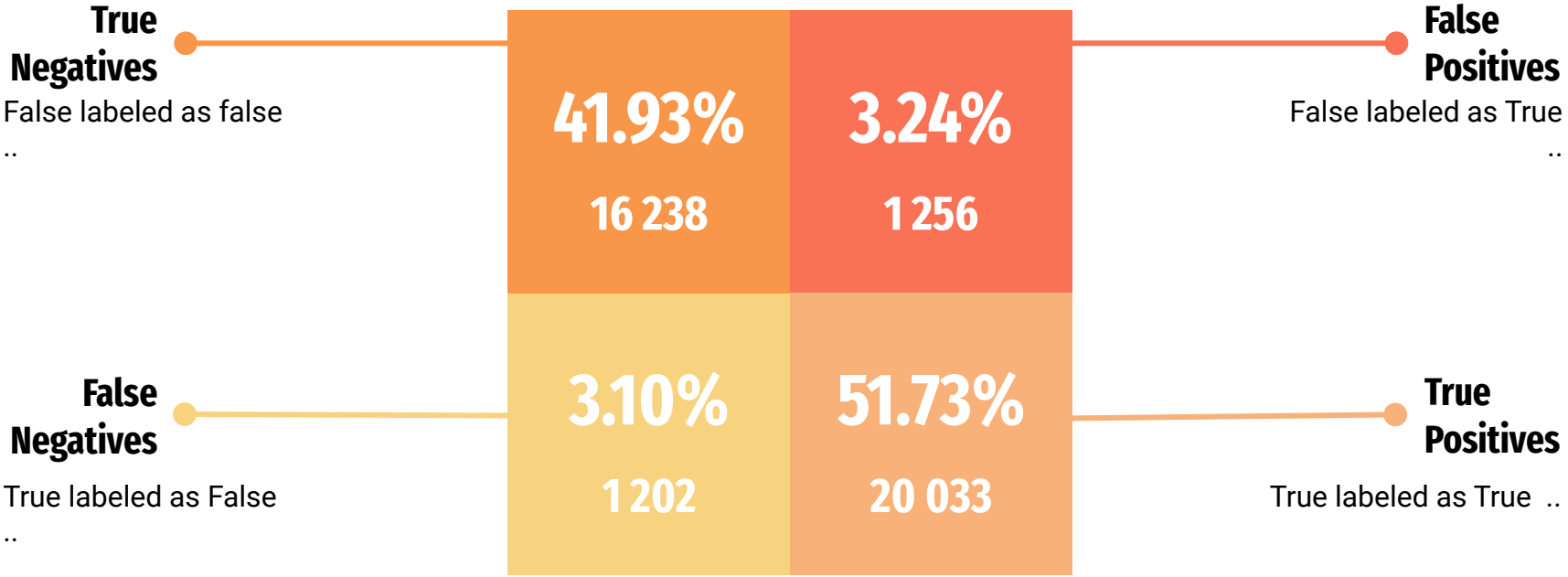
What proportion of actual negatives was identified correctly?

93.5%

AUC SCORE

How much the model is capable of distinguishing between classes

Confusion Matrix



Thanks for your Attention!

Any Questions?