



Tableau – The Cathy Airlines

Field: Business Intelligence

Topic: Tableau and WEKA

University: Middlesex University

Author: Agata Gabara (M00728162)

Supervisor: Geili ElSanousi

Date: April 2024

Introduction.....	3
Data analysis and Visualisation – Tableau.....	10
The pie chart by gender.....	10
Departure delay by class and type of travel.....	11
Flight distance by customer and gender	13
Tree map by class group.....	14
Average departure delay vs. arrival delay by age.....	16
Average flight distance vs. average delay by age	18
Calculations LoD	20
Dashboard	23
Comparison of classes: Business, Eco and Eco Plus	23
Distribution of Age for Men and Women.....	25
Arrival delay in minutes and arrival delay in minutes -dual combination.....	26
Cluster analysis of flight delays.....	28
WEKA – k-means algorithm	30
Data ethics.....	36
Conclusions.....	37

Introduction

The data set has been downloaded from website:

www.kaggle.com. (n.d.). *Airlines Customer satisfaction*. [online] Available at:

https://www.kaggle.com/datasets/sjleshtrac/airlines-customer-satisfaction?select=Invistico_Airline.csv

It is entitled "Airlines Customer Satisfaction". The author is Sayantan Jana.

The dataset is made up of the characteristics of customers who have already used the line. The opinion of the consumers on different context and their flight data has been merged. The main goal of this dataset is to forecast if a future client would be happy with their service given the details of the other variables values. Moreover, the Cathy Airlines would like to know which aspect of the services given by them have to be highlighted more to produce more satisfied customers.

Original data set - characteristics: bigger

- 129880 rows (1 lines as header, 129881-1)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
129867	satisfied	Female	disloyal C	59	Personal T	Eco	2641	4	5	4	3	2	4	2	2
129868	satisfied	Female	disloyal C	18	Personal T	Eco	1243	5	0	5	1	2	5	2	2
129869	satisfied	Female	disloyal C	30	Personal T	Eco	1961	5	1	5	4	5	5	5	5
129870	satisfied	Female	disloyal C	45	Personal T	Eco	1612	5	2	5	3	2	5	2	2
129871	satisfied	Female	disloyal C	55	Personal T	Eco	1953	5	2	5	4	1	5	5	1
129872	satisfied	Female	disloyal C	70	Personal T	Eco	1674	5	4	5	1	5	5	5	5
129873	satisfied	Female	disloyal C	35	Personal T	Eco	3287	5	4	5	3	2	5	2	2
129874	satisfied	Female	disloyal C	69	Personal T	Eco	2240	5	4	5	3	4	5	4	4
129875	satisfied	Female	disloyal C	63	Personal T	Eco	1942	5	5	4	4	3	4	3	3
129876	satisfied	Female	disloyal C	11	Personal T	Eco	2752	5	5	5	2	2	5	2	2
129877	satisfied	Female	disloyal C	29	Personal T	Eco	1731	5	5	5	3	2	5	2	2
129878	dissatisfie	Male	disloyal C	63	Personal T	Business	2087	2	3	2	4	2	1	1	3
129879	dissatisfie	Male	disloyal C	69	Personal T	Eco	2320	3	0	3	3	3	2	2	4
129880	dissatisfie	Male	disloyal C	66	Personal T	Eco	2450	3	2	3	2	3	2	2	3
129881	dissatisfie	Female	disloyal C	38	Personal T	Eco	4307	3	4	3	3	3	3	3	4

P	Q	R	S	T	U	V	W
1	5	2	3	3	2	5	0
4	5	5	3	5	2	0	0
3	2	3	4	4	5	0	0
2	5	4	3	4	2	0	0
1	1	3	3	4	1	0	0
3	2	4	5	4	5	54	46
4	5	4	4	3	2	9	0
5	4	4	3	4	4	4	0
5	2	5	3	5	3	7	
3	5	3	5	4	2	5	0
3	3	4	4	4	2	0	0
2	3	3	1	2	1	174	172
4	3	4	2	3	2	155	163
3	2	3	2	1	2	193	205
5	5	5	3	3	3	185	186

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	
129867	satisfied	Female	disloyal	Ci	59	Personal	TEco	2641	4	5	4	3	2	4	2	2	1	5	2	3	3	2	5	0
129868	satisfied	Female	disloyal	Ci	18	Personal	TEco	1243	5	0	5	1	2	5	2	2	4	5	5	3	5	2	0	0
129869	satisfied	Female	disloyal	Ci	30	Personal	TEco	1961	5	1	5	4	5	5	5	5	3	2	3	4	4	5	0	0
129870	satisfied	Female	disloyal	Ci	45	Personal	TEco	1612	5	2	5	3	2	5	2	2	2	5	4	3	4	2	0	0
129871	satisfied	Female	disloyal	Ci	55	Personal	TEco	1953	5	2	5	4	1	5	5	1	1	1	3	3	4	1	0	0
129872	satisfied	Female	disloyal	Ci	70	Personal	TEco	1674	5	4	5	1	5	5	5	5	3	2	4	5	4	5	54	46
129873	satisfied	Female	disloyal	Ci	35	Personal	TEco	3287	5	4	5	3	2	5	2	2	4	5	4	4	3	2	9	0
129874	satisfied	Female	disloyal	Ci	69	Personal	TEco	2240	5	4	5	3	4	5	4	4	5	4	4	3	4	4	4	0
129875	satisfied	Female	disloyal	Ci	63	Personal	TEco	1942	5	5	4	4	3	4	3	3	5	2	5	3	5	3	7	
129876	satisfied	Female	disloyal	Ci	11	Personal	TEco	2752	5	5	5	2	2	5	2	2	3	5	3	5	4	2	5	0
129877	satisfied	Female	disloyal	Ci	29	Personal	TEco	1731	5	5	5	3	2	5	2	2	3	3	4	4	4	2	0	0
129878	dissatisfie	Male	disloyal	Ci	63	Personal	TBusiness	2087	2	3	2	4	2	1	1	3	2	3	3	1	2	1	174	172
129879	dissatisfie	Male	disloyal	Ci	69	Personal	TEco	2320	3	0	3	3	3	2	2	4	4	3	4	2	3	2	155	163
129880	dissatisfie	Male	disloyal	Ci	66	Personal	TEco	2450	3	2	3	2	3	2	2	3	3	2	3	2	1	2	193	205
129881	dissatisfie	Female	disloyal	Ci	38	Personal	TEco	4307	3	4	3	3	3	3	3	4	5	5	5	3	3	3	185	186

- 23 columns

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	satisfactio	Gender	Customer	Age	Type of Tra	Class	Flight Dist	Seat comf	Departure	Food and	Gate locat	Inflight wif	Inflight ent	Online sup	Ease of Or	On-board	Leg room	Baggage h	Checkin sr	Cleanlines	Online bo	Departure	Arrival Delay in Minutes		

- text and numeric values

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	satisfied	Gender	Customer	Age	Type of Tra	Class	Flight Dist	Seat comf	Departure	Food and	Gate locat	Inflight wif	Inflight ent	Online sup	Ease of Or	On-board	Leg room	Baggage h	Checkin s	Cleanlines	Online bo	Departure	Arrival	Delay in Minutes
2	satisfied	Female	Loyal Cust	65	Personal T	Eco	265	0	0	0	2	2	4	2	3	3	0	3	5	3	2	0	0	
3	satisfied	Male	Loyal Cust	47	Personal T	Business	2464	0	0	0	3	0	2	2	3	4	4	4	2	3	2	310	305	
4	satisfied	Female	Loyal Cust	15	Personal T	Eco	2138	0	0	0	3	2	0	2	2	3	3	4	4	4	2	0	0	
5	satisfied	Female	Loyal Cust	60	Personal T	Eco	623	0	0	0	3	3	4	3	1	1	0	1	4	1	3	0	0	
6	satisfied	Female	Loyal Cust	70	Personal T	Eco	354	0	0	0	3	4	3	4	2	2	0	2	4	2	5	0	0	
7	satisfied	Male	Loyal Cust	30	Personal T	Eco	1894	0	0	0	3	2	0	2	2	5	4	5	5	4	2	0	0	
8	satisfied	Female	Loyal Cust	66	Personal T	Eco	227	0	0	0	3	2	5	5	5	5	0	5	5	5	3	17	15	
9	satisfied	Male	Loyal Cust	10	Personal T	Eco	1812	0	0	0	3	2	0	2	2	3	3	4	5	4	2	0	0	
10	satisfied	Female	Loyal Cust	56	Personal T	Business	73	0	0	0	3	5	3	5	4	4	0	1	5	4	4	0	0	
11	satisfied	Male	Loyal Cust	22	Personal T	Eco	1556	0	0	0	3	2	0	2	2	2	4	5	3	4	2	30	26	
12	satisfied	Female	Loyal Cust	58	Personal T	Eco	104	0	0	0	3	3	3	3	3	3	0	1	2	3	5	47	48	
13	satisfied	Female	Loyal Cust	34	Personal T	Eco	3633	0	0	0	4	2	0	2	2	3	2	5	2	5	2	0	0	
14	satisfied	Male	Loyal Cust	62	Personal T	Eco	1695	0	0	0	4	5	0	5	5	1	3	2	2	4	5	0	0	
15	satisfied	Male	Loyal Cust	35	Personal T	Eco	1766	0	1	0	1	4	0	4	4	3	5	2	3	2	4	0	0	
16	satisfied	Female	Loyal Cust	47	Personal T	Eco	84	0	1	0	1	5	2	1	5	5	0	5	2	5	2	40	48	
17	satisfied	Male	Loyal Cust	60	Personal T	Eco	1373	0	1	0	1	1	0	1	1	3	4	1	4	2	1	0	0	
18	satisfied	Female	Loyal Cust	13	Personal T	Eco	3693	0	1	0	2	4	0	4	4	4	4	1	3	1	4	5	0	
19	satisfied	Female	Loyal Cust	52	Personal T	Business	2610	0	1	0	2	1	2	2	1	1	0	1	2	1	3	0	0	
20	satisfied	Female	Loyal Cust	55	Personal T	Eco	2554	0	1	0	2	0	1	1	2	1	1	2	1	3	1	0	0	
21	satisfied	Female	Loyal Cust	28	Personal T	Eco	3095	0	1	0	2	3	0	3	3	2	5	2	3	2	3	0	0	
22	satisfied	Female	Loyal Cust	9	Personal T	Eco	3305	0	1	0	2	3	0	5	3	1	1	1	3	3	3	0	0	
23	satisfied	Female	Loyal Cust	10	Personal T	Eco	2090	0	1	0	2	1	0	1	1	3	5	1	4	2	1	0	0	
24	satisfied	Female	Loyal Cust	25	Personal T	Eco	2122	0	1	0	2	2	0	4	2	4	1	3	1	3	2	0	0	
25	satisfied	Male	Loyal Cust	53	Personal T	Business	1099	0	1	0	2	1	3	3	1	1	0	1	3	1	1	0	0	
26	satisfied	Female	Loyal Cust	16	Personal T	Eco Plus	1747	0	1	0	2	2	0	2	2	3	3	2	4	3	2	0	0	
27	satisfied	Male	Loyal Cust	30	Personal T	Eco	1817	0	1	0	2	4	0	4	4	2	1	3	3	2	4	0	0	
28	satisfied	Male	Loyal Cust	64	Personal T	Eco	1707	0	1	0	2	5	0	3	5	4	4	2	3	2	5	0	0	
29	satisfied	Female	Loyal Cust	42	Personal T	Eco	470	0	1	0	2	3	2	2	3	3	0	3	1	3	4	2	23	

Cleaning – activities:

- Firstly, 128 000 rows have been deleted by the usage of function “RAND”. The function has brought each row the random value and after that I have sorted in descending order, and I have chosen the first 1010 rows (others- have been deleted),
- Secondly, by the usage of function COUNTBLANK, I have checked how many fields have not got any values; 3 fields have been empty (column W). In the empty fields, I have filled “0”. After this operation, I have data continuity,

Function COUNTBLANK

=COUNTBLANK(W2:W1010)

Range: (W2:W1010)

Location of empty cells:

	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
742	IT Eco	1636	3	2	2	3	2	2	1	2	3	4	4	2	4	2	0	0
743	st Business	382	4	1	4	4	3	4	5	5	5	5	5	3	5	4	10	2
744	IT Eco	1899	4	4	4	4	2	5	5	5	5	5	5	5	5	1	0	0
745	st Business	2901	2	2	2	2	2	4	5	5	5	5	5	5	5	3	0	9
746	st Eco	2877	2	5	1	1	2	2	2	2	4	3	4	3	4	2	0	0
747	st Business	1571	4	3	3	3	4	3	3	4	4	4	4	1	4	1	0	0
748	st Business	873	3	3	3	3	2	5	4	4	4	4	4	3	4	4	4	4
749	IT Eco	1849	5	5	5	5	2	5	4	5	5	5	5	5	5	3	0	0
750	st Business	1394	3	3	3	3	3	3	3	3	4	5	5	3	5	3	102	
751	IT Eco	1463	4	1	4	3	3	4	3	3	4	4	4	4	3	3	8	2
752	IT Eco	3432	3	3	3	3	2	5	5	4	4	4	4	4	4	3	0	0
753	IT Eco	2465	2	4	2	3	2	2	2	2	4	2	5	4	4	2	5	0
754	st Eco	335	1	5	5	5	5	3	4	1	1	1	1	4	1	4	46	40
755	st Eco	1708	4	0	4	5	2	4	2	2	2	4	4	2	3	2	0	0
756	st Eco	2397	2	2	2	4	1	2	1	1	5	5	5	3	2	1	4	0
757	st Business	1105	1	1	1	1	2	5	4	3	3	3	3	4	3	5	20	7
758	st Eco	1920	4	2	2	2	4	4	4	4	1	1	3	2	3	4	0	0
759	st Eco	1757	1	4	4	4	1	1	1	1	2	3	4	3	4	1	0	0
760	st Business	1405	1	1	1	1	3	5	5	4	4	4	4	5	4	5	3	0

(W750)

	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
928	st Eco	1789	1	3	0	3	5	0	5	5	3	3	4	2	4	5	62	58
929	st Business	873	2	5	5	5	5	3	4	2	2	2	2	4	2	1	6	9
930	st Business	2650	2	2	2	2	3	4	3	2	2	2	2	2	2	1	26	28
931	st Eco Plus	387	2	4	5	5	4	4	3	3	3	2	3	4	3	4	0	0
932	st Eco	1853	2	1	1	1	2	3	2	2	4	2	3	4	3	2	49	58
933	st Business	947	3	3	3	3	4	5	4	5	5	5	5	4	5	4	0	5
934	st Business	714	1	1	1	1	4	4	5	3	3	3	3	5	3	4	0	
935	st Business	907	1	1	1	1	4	5	5	4	4	4	4	5	4	4	0	0

(W934)

	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
231	st Business	2789	2	2	2	4	5	2	5	5	4	2	5	5	4	5	0	0
232	IT Eco Plus	1257	3	4	4	5	4	4	4	4	3	4	4	3	5	4	11	6
233	IT Eco	1832	3	4	3	5	4	3	4	4	3	2	5	3	4	4	0	0
234	st Business	3464	1	1	1	1	4	4	4	5	5	5	5	4	5	5	0	0
235	IT Eco	1237	2	5	2	1	3	2	3	3	5	3	4	4	4	3	0	6
236	IT Eco	1542	2	4	5	2	2	5	1	2	4	4	4	3	4	2	0	0
237	IT Eco	2047	1	2	1	3	2	1	2	2	2	5	4	4	3	2	0	0
238	st Eco Plus	744	1	4	4	4	5	4	3	1	1	1	1	4	1	4	9	6
239	st Eco	2373	4	4	4	4	5	4	5	5	3	5	4	4	5	5	0	0
240	st Business	3385	3	3	3	3	3	3	3	3	4	2	3	2	4	3	0	7
241	IT Eco	2259	1	1	1	1	4	4	4	4	5	2	4	5	4	4	0	3
242	st Eco	3486	1	1	1	2	2	1	3	2	2	2	1	1	2	2	0	0
243	st Business	3183	4	4	4	4	5	5	5	5	5	5	5	3	5	4	0	2
244	st Business	3110	4	2	4	4	4	5	4	5	5	5	5	4	5	5	0	0
245	st Business	2393	2	2	3	2	5	5	5	5	5	5	5	3	5	4	3	

(W245)

- There are 14 categorical values (values: 0-5) concerning customer satisfaction, I have eliminated 6 columns where some values are missing, for example “0”. I have been not interested in columns such as: gate location, online support, on board service, baggage handling, check in service and cleanliness,

H	I	J	K	L	M	N	O	P
Seat comfort	Departure	Food and c	Ease of On	Leg room s	Online boa	Departure	Arrival Delay in Minutes	
4	3	3	4	2	4	2	9	
4	4	4	4	4	4	0	0	
1	2	1	1	1	1	0	44	
3	3	3	4	5	5	5	19	
5	5	5	5	1	5	5	27	
0	0	0	4	4	5	0	0	
1	1	1	4	1	4	0	8	
3	3	4	3	3	3	0	12	
1	1	1	3	3	3	109	108	
2	5	2	1	5	1	35	29	
1	0	0	1	0	3	14	11	
4	4	4	5	4	3	0	0	
4	4	4	4	5	4	0	0	
2	3	4	2	2	2	186	172	
4	4	4	5	5	4	42	44	
1	1	1	4	4	4	0	0	
2	4	4	3	3	2	133	157	
0	0	0	1	4	1	0	0	
3	2	3	5	1	5	0	0	

- For the columns which I have left, I have done mapping; there have been values from 0 to 5 and they have had verbal descriptions such as:

0	Highly dissatisfied
1	Somewhat dissatisfied
2	Neutral
3	Satisfied with room for improvement
4	Very satisfied with minor suggestions
5	Extremely satisfied, no reservations

*This table has been existed in Excel file as the new bookmark.

- I have used Excel function VLOOKUP. For column "Seat comfort" and numerical values in it,

H
Seat comfort
4
4
1
3
5
0
1
3
1
2
1
4
4
2
4
1

I have brought word description to each value from 0-5 as below.

H
Seat comfort
Very satisfied with minor suggestions
Very satisfied with minor suggestions
Somewhat dissatisfied
Satisfied with room for improvement
Extremely satisfied, no reservations
Highly dissatisfied
Somewhat dissatisfied
Satisfied with room for improvement
Somewhat dissatisfied
Neutral
Somewhat dissatisfied
Very satisfied with minor suggestions
Very satisfied with minor suggestions
Neutral
Very satisfied with minor suggestions
Somewhat dissatisfied

*At the end, I have pasted as values what I have received as it will be not possible use the values with formula VLOOKUP.

* This activity has been done for all categorical columns.

- The columns: K and L have been deleted as they have not brought anything new to analysis.

Final version

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	satisfaction	Gender	Customer	Age	Type of Tra	Class	Flight Dist	Seat comf	Departure, Food and c	Ease of On	Leg room s	Online bo	Departure	Arrival	Delay in	Minutes	

16 columns

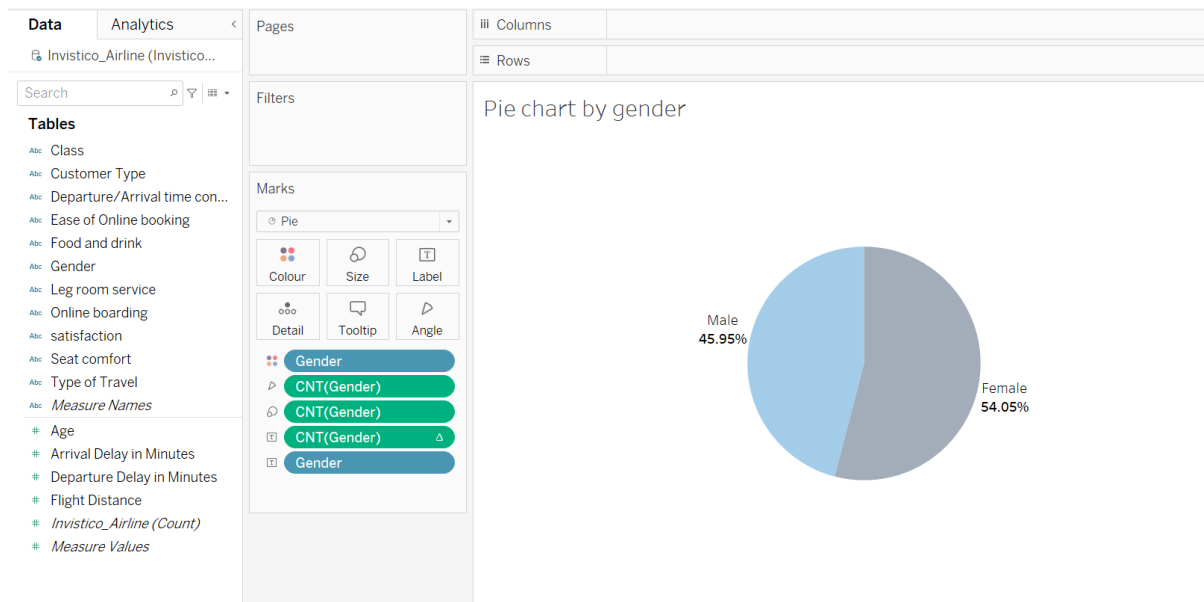
1010 rows

1002
1003
1004
1005
1006
1007
1008
1009
1010
1011

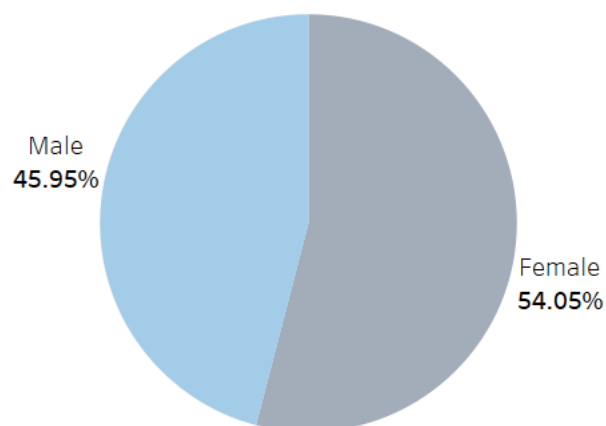
Data analysis and Visualisation

1.The pie chart by gender.

- I choose pie chart because I would like to see what the percentage of males and females is in the whole group. I don't need any scale. I need a simple visual representation divided on slices.



Pie chart by gender

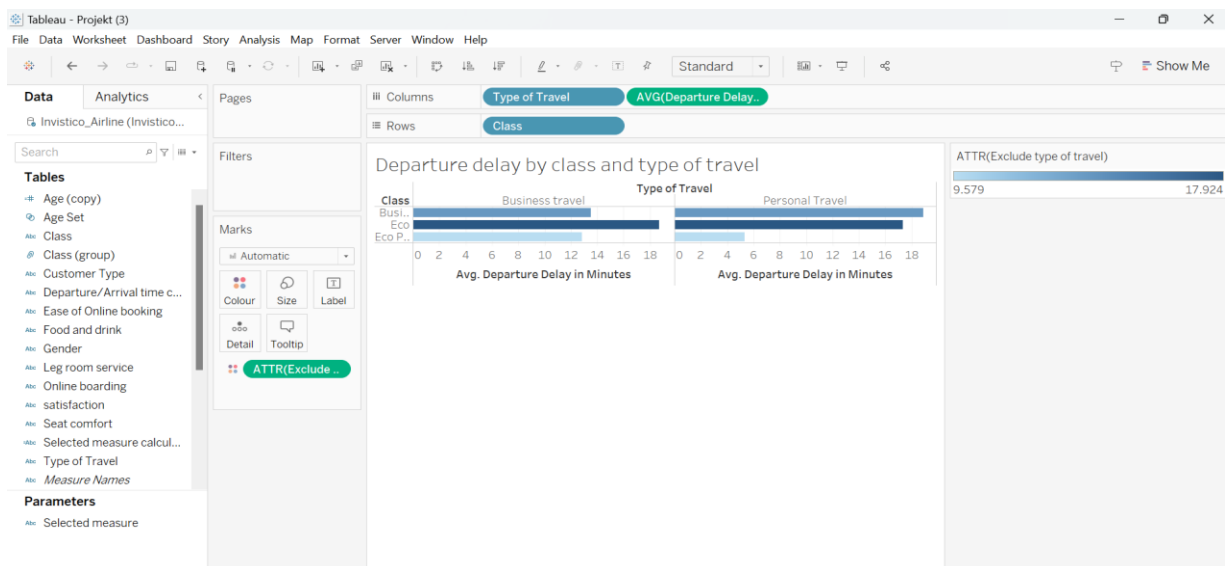


Description of the graph

The graph is entitled of “ The pie chart by gender”. It presents the percentage of men and women in the overall sample. There is a small difference between men and women; 8.1%. The female share is slightly higher than male one.

2. Departure delay by class and type of travel.

- I select horizontal bars to present 2 dependencies: average departure delay in minutes and respective classes (business, eco, eco plus). There are 2 axes: horizontal and vertical. On horizontal, there is an average delay presented in minutes. Each class has different average delay time, different length of horizontal bar. To create this graph, I need dimension (rows) and measure (columns). Dimension is Class, Measure is Average departure time in minutes.

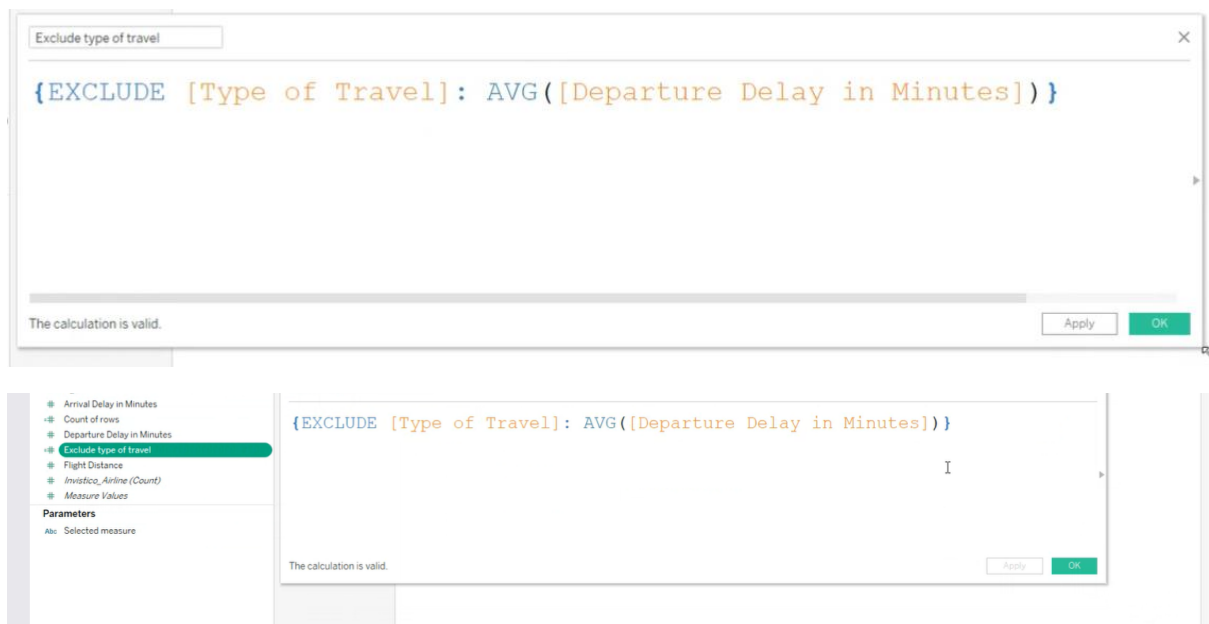



Sorting by axis

I have used EXCLUDE to get independent colour palette for different type of travel.

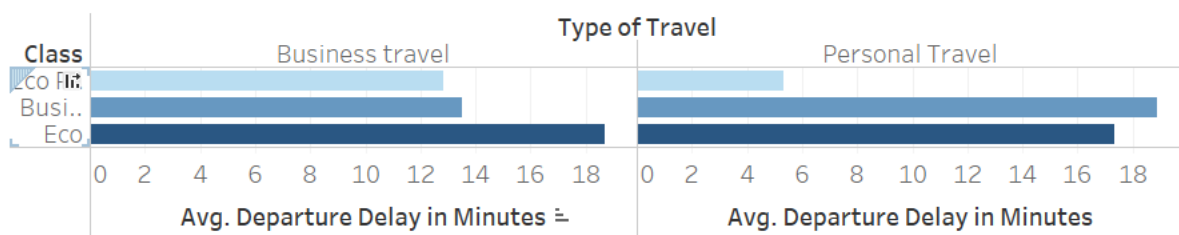
I have 1 scale for both graphs, I would like to have 2 scales for 2 graphs.



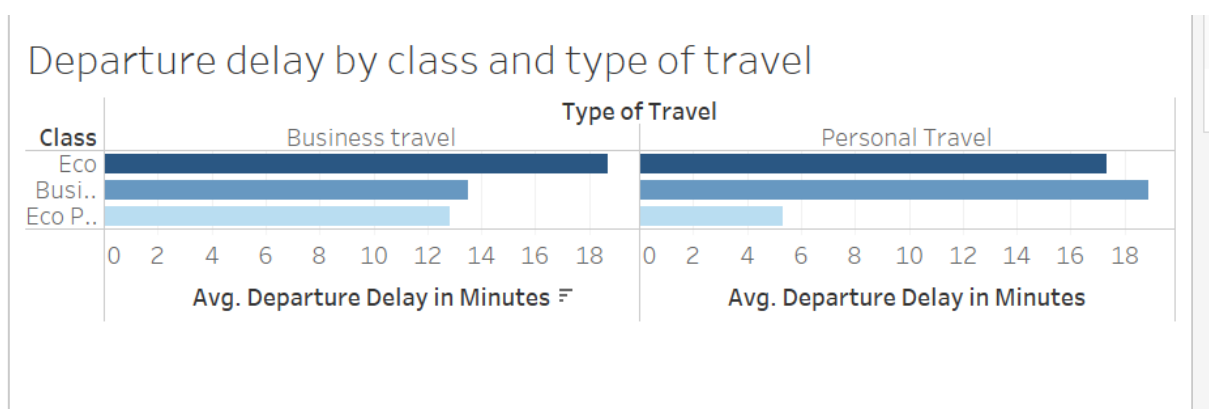


Sorting by axis is automatically; I can click this icon  and sorting.

Departure delay by class and type of travel



Or



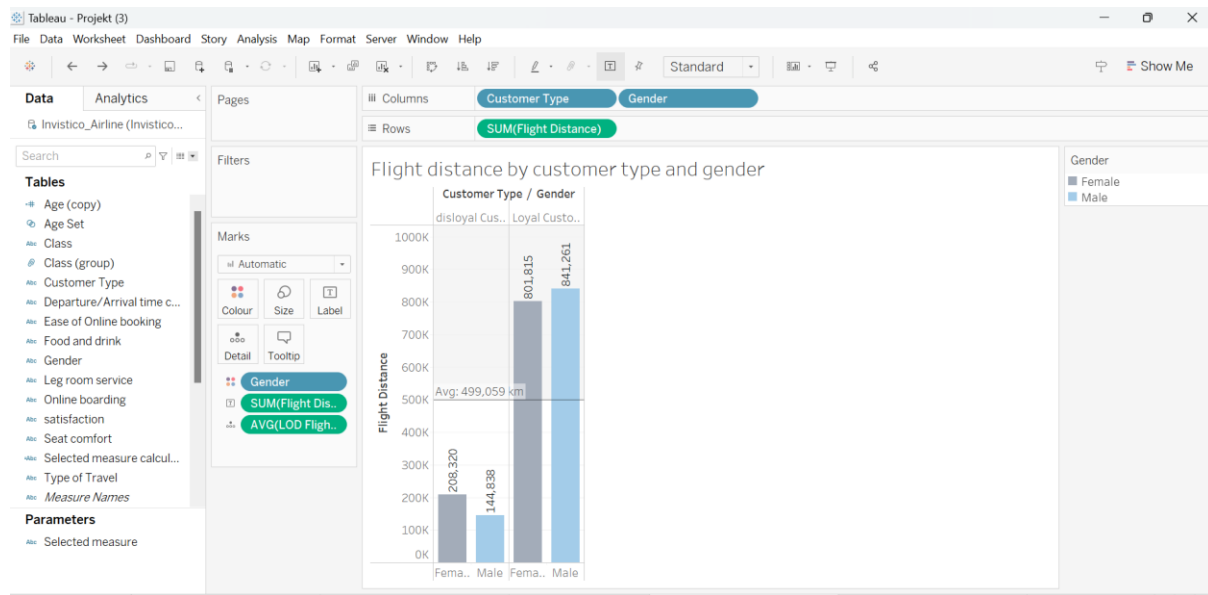
*Sorting by toolbar is not automatically.

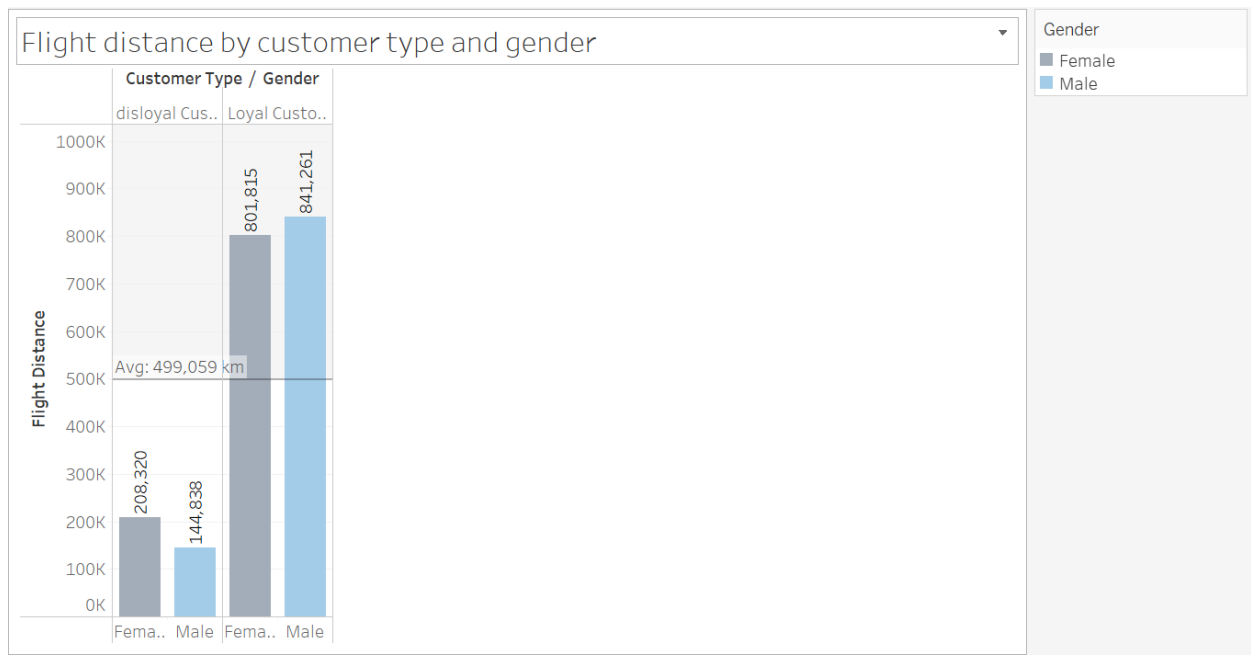
Description of the graph

The graph is entitled of “Departure delay by class and type of travel”. There are 3 types of class: Business, Eco and Eco Plus. There are 2 types of travel: Business and Personal. On the x-axis it is shown an average departure delay in minutes for respective type of travel. The highest delay (18.68 minutes) in business travel is in Eco class. Between Business and Eco Plus class the average departure delay doesn’t differ significantly (13.51, 12.84). In Personal Travel, the smallest average delay in minutes is in Eco Plus class (5.33), between Business and Eco there is small difference; (18.9-17.32=1.58 min).

3. Flight distance by customer and gender.

- I have chosen bar chart because I would like to compare values over various categories.





Description of the graph

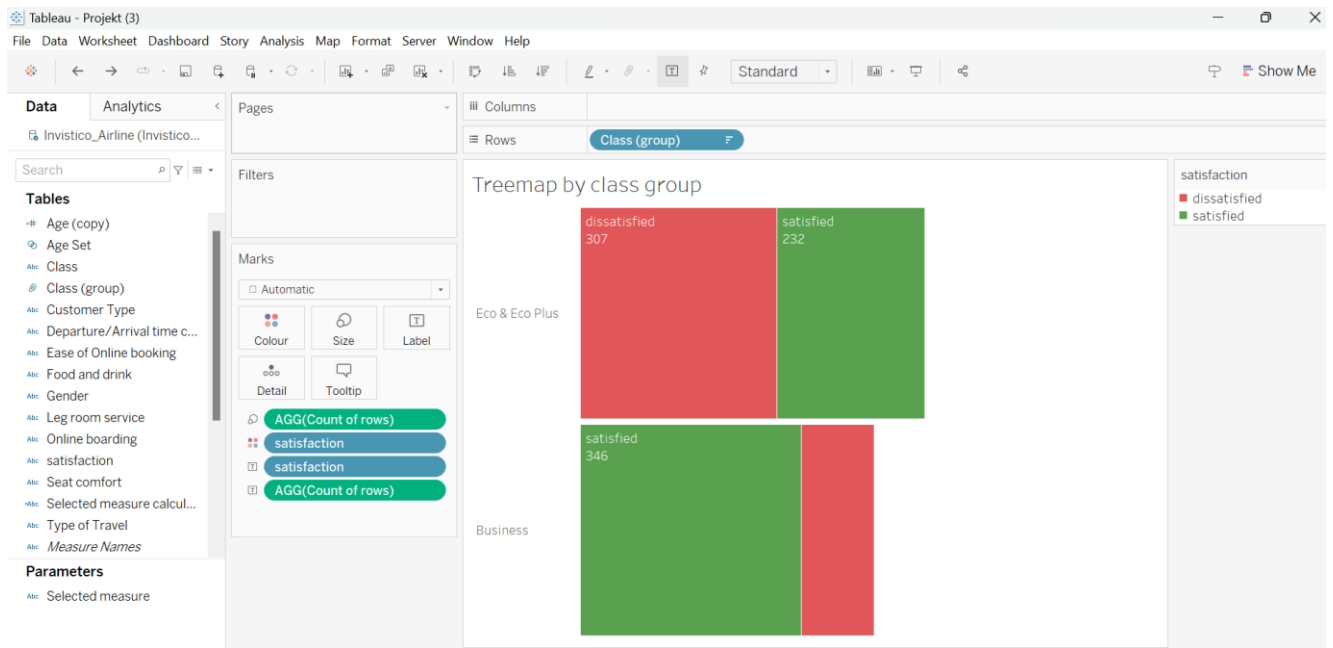
The chart illustrates the flight distance by customers type and gender. On the x-axis, gender is depicted while they y-axis represents flight distance in kilometres. The illustration is divided into two distinct sections: disloyal and loyal customers. In the left section of the graph, disloyal customers are depicted. Females are categorized a disloyal customer as at a flight distance of 208,320 kilometres, while males are considered disloyal at a flight distance of 144,838 kilometres. Moving to the right section of the presentation, the loyal clientele with flight distances exceeding 800,000 kilometres is showcased. Males are recognised as loyal customers at a distance of 841,261 kilometres, whereas females are considered loyal at distance of 801,815 kilometres. Additionally, a reference line is set at 499,059 kilometres for comparison purposes. In the top right corner of illustration, a legend is provided for clarity regarding the representation of different customer types and genders.

AVG LOD Flight Distance = Flight Distance.

4. Tree map by class group.

- I decide to use a tree map because I would like to indicate hierarchical data in nested rectangles. Every rectangle presents a hierarchical level, and its size correlates with metric, frequency, value or size. The hierarchy is enacted by the nesting of rectangles within one another, with the peripheral rectangle presenting the highest level of the hierarchy and inside rectangle showing lower level of the hierarchy.
- I use dimensions to describe the skeleton of the tree map and measure to identify the colour and size of the individual rectangles. Colour: Satisfaction

(satisfied or dissatisfied), Size: AGG (Count of rows), Label: Satisfaction, AGG (Count of rows).



Treemap by class group



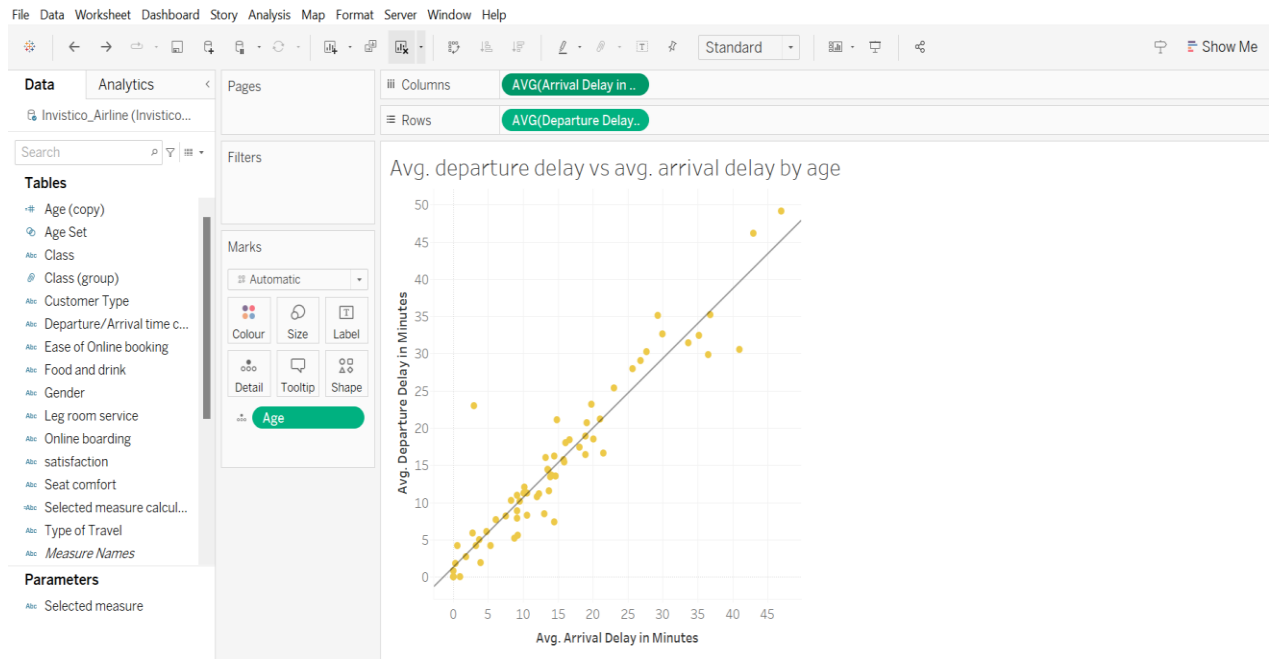
Description of the graph

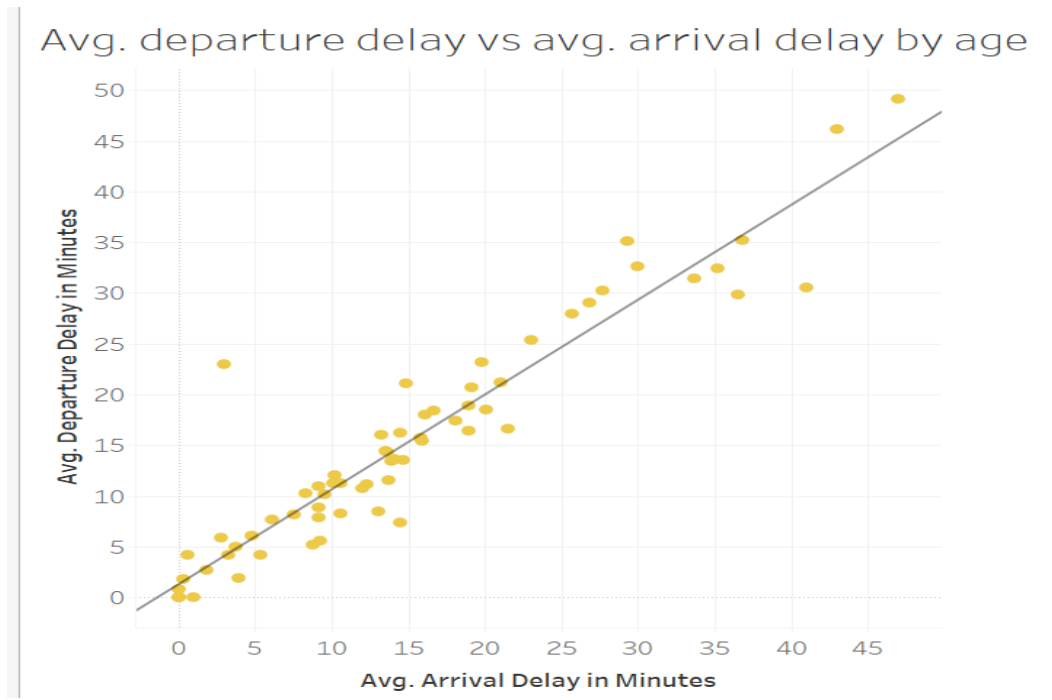
The illustration titled “Tree map by class group” presents 3 distinct group: Eco, Eco Plus and Business. Notably, Eco and Eco Plus have been consolidated into a single entity. Within each group, 2 categories are represented: dissatisfied and satisfied.

In the combined Eco and Eco Plus, the dissatisfied segment outweighs the satisfied segment, with a notable difference of 75 individuals (307 dissatisfied versus 232 satisfied). Conversely, within the Business group, the satisfied category predominates, with 232 individuals expressing satisfaction.

5. Average departure delay vs. arrival delay by age.

- I opt for scatter plot because I have 2 numeric fields.





Description of the graph

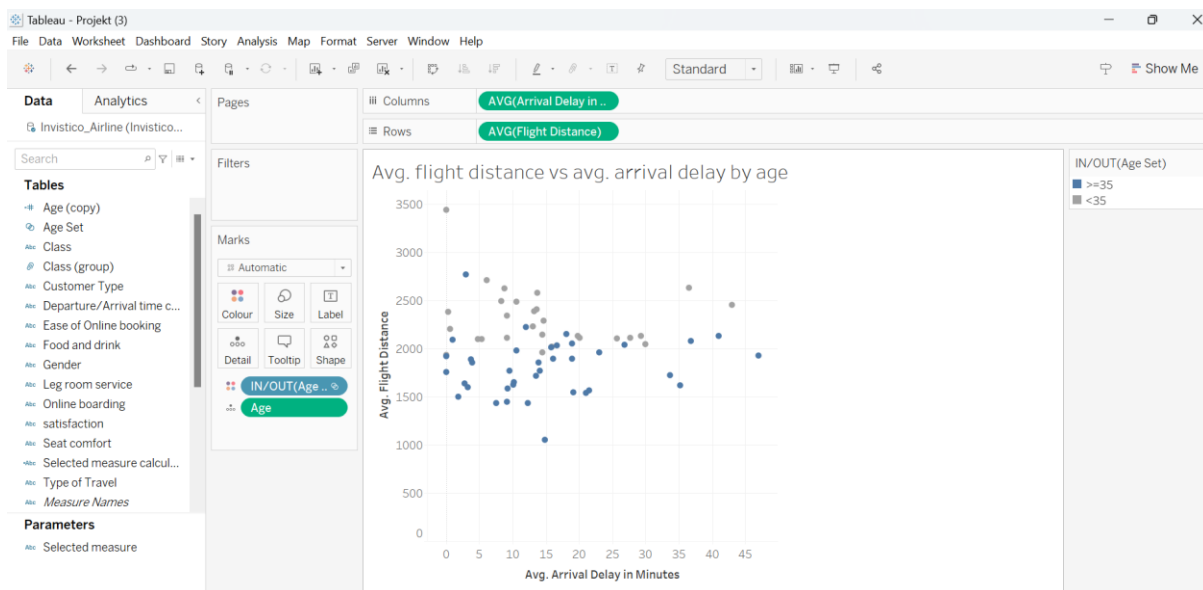
The illustration titled “Average Departure Delay versus Average Arrival Delay by Age” reveals a notable dependency between two key averages: the average departure delay in minutes and the average arrival delay in minutes. This dependency is visually represented by a linear trend line, providing insight into the data distribution.

With an R-squared value of 0.89268, the linear fit of the trend line is strong, indicating a substantial portion of the variance in arrival delay can be explained by departure delay. The p-value, being less than 0.0001, underscores the statistical significance of this relationship. Notably, the R-squared value, though not perfect reflects a robust fit, while the small p-value confirms the reliability of the observed relationship. Some data points fall above the trend line, signifying instances where arrival delays exceed departure delays, while others fall below it, suggesting instances of shorter arrival delays compared to departure delays.

*



Avg. Departure Delay in Minutes = $0.934739 \times \text{Avg. Arrival Delay in Minutes} + 1.28582$
R-Squared: 0.89268
P-value: < 0.0001

6. Average flight distance vs. average delay by age.



Discreet values (on these values it is possible to create a set),

Tables

-  Age Set
- Abc Class
-  Class (group)
- Abc Customer Type
- Abc Departure/Arrival time c...
- Abc Ease of Online booking
- Abc Food and drink
- Abc Gender
- Abc Leg room service
- Abc Online boarding
- Abc satisfaction
- Abc Seat comfort
- Abc Selected measure calcul...
- Abc Type of Travel
- Abc *Measure Names*

Copying set Age,

Age(copy) transfer to dimensions,

Creating conditional set,

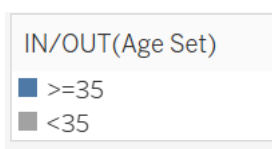
Description of the graph

The illustration titled “Average flight distance versus Average Arrival Delay by Age” offers insights into the relationship between average arrival delay in minutes and average flight distance, with age as conditioning factor. On the x-axis, the average arrival delay is plotted, while the y-axis represents the average flight distance.

Distinct groups are delineated by colour: individuals aged 35 and older are depicted in blue, while those younger than 35 are shown in grey. Notably, individuals aged 35 and older tend to dominate the average flight distance range, spanning from 1437km to 2224km. Conversely, customers younger than 35 years old exhibit dominance in the distance range of 1960 km to 2711km.

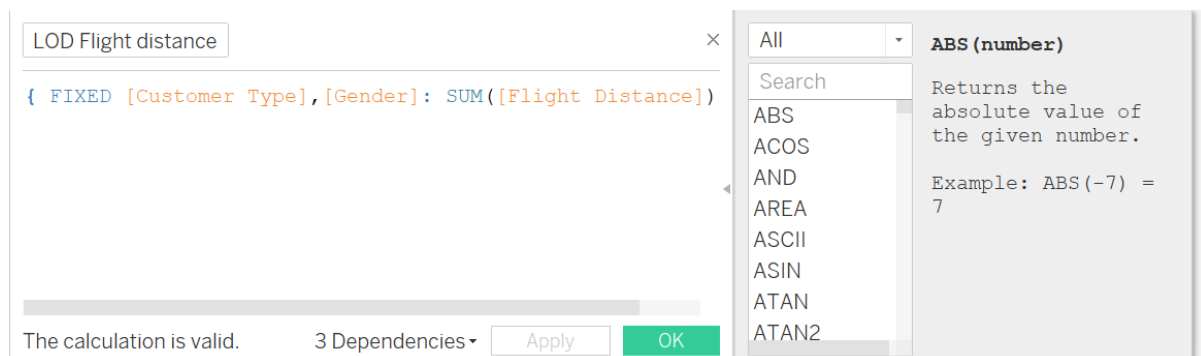
Across both age groups, the average arrival delay fluctuates between 0 minutes and 47 minutes, indicating variability in punctuality regardless of age.

*



7.Calculations LoD

LOD Flight distance - formula



- LoD shows on each level that data is aggregated.
- I have created additional table to show how LoD works.

Edit Reference Line, Band, or Box

Line Band Distribution Box Plot

Scope

☒ Entire Table ☐ Per Pane ☐ Per Cell

Line

Value: AVG(LOD Flight distance) Average

Label: Custom Avg: <Value> km

Tooltip: None

Line only 95

Formatting

Line: Fill Above: Fill Below: None

☐ Show recalculated line for highlighted or selected data points

OK

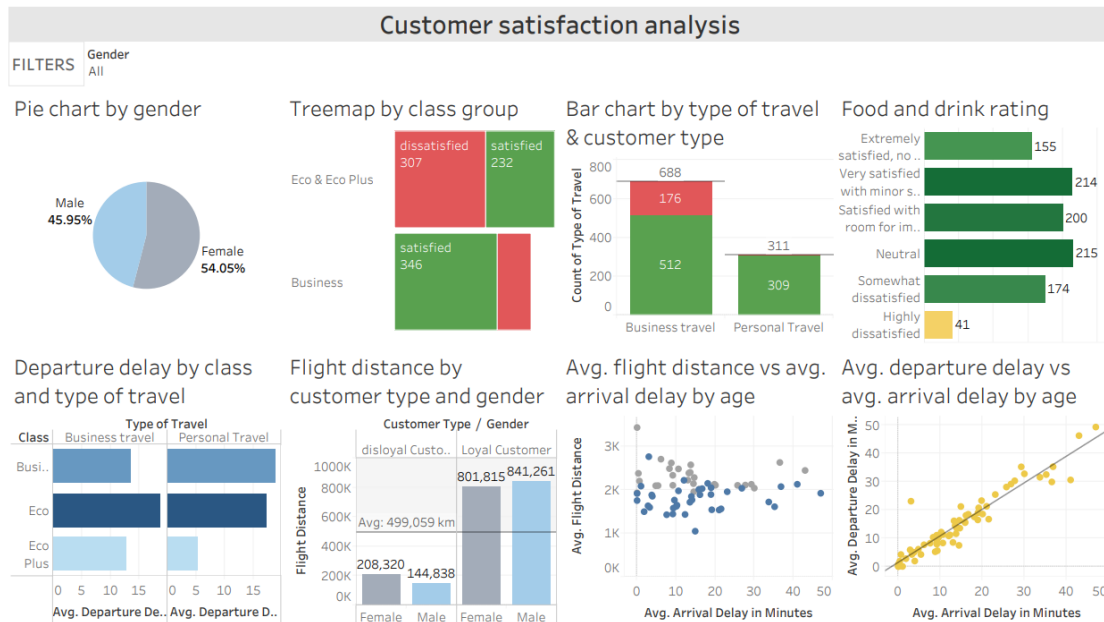
Description

The graph illustrates a segmentation based on customer type and gender, providing insights into the flying behaviour of each demographic. Utilising Lo (level of detail) calculations, the graph displays the average flight distance for each customer group while also highlighting the distribution across gender.

8. Dashboard

Summary

- 8 graphs put on the dashboard and adjust accordingly.

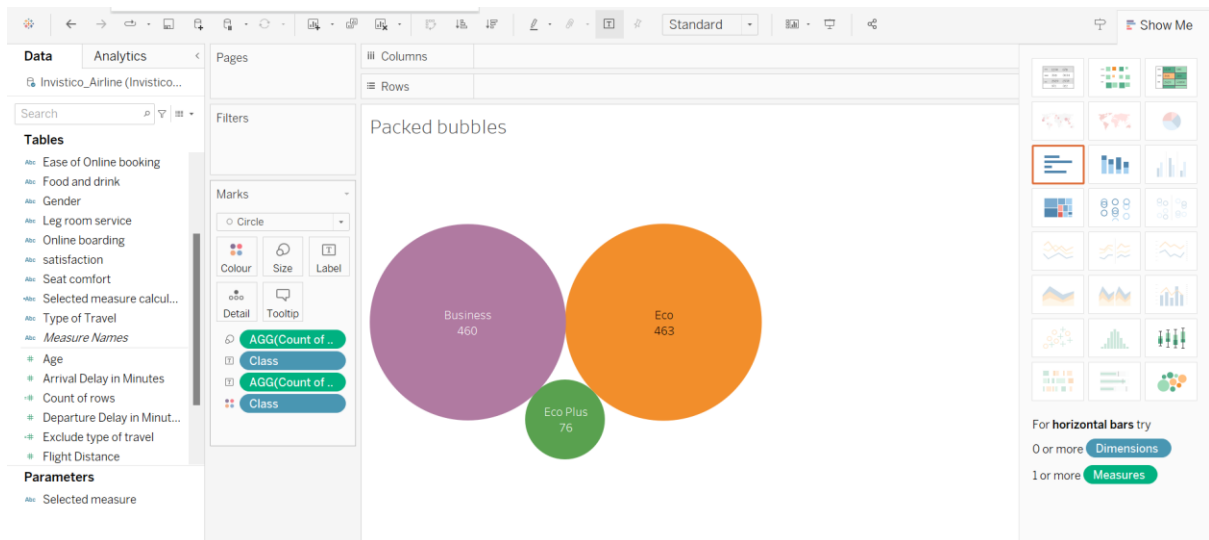


Description of graph

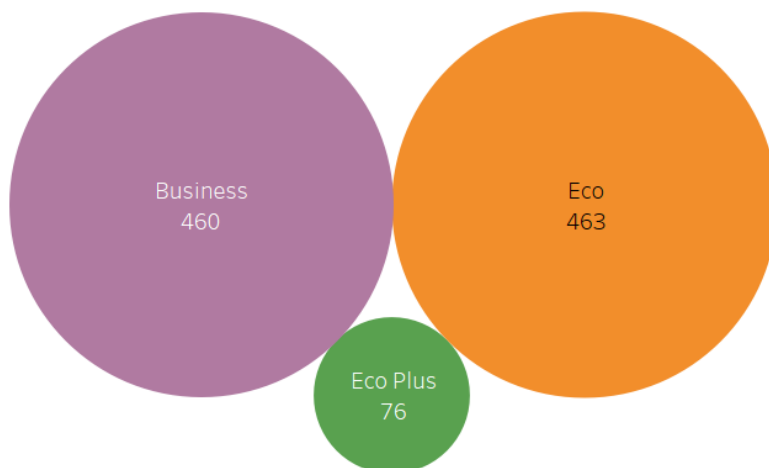
This dashboard is made up of 8 graphs: (Pie chart by gender, Tree map by class group, Bar chart by type of travel & customer type, Food and drink rating, Departure delay by class and type of travel, Flight distance by customer type and gender, Average flight distance vs. avg. arrival delay by age, Avg. departure delay vs. avg. arrival delay by age).

9. Comparison of classes: Business, Eco and Eco Plus.

This graph shows data in circles' cluster. This type of graph it is used when there are at least 3 variables to compare.



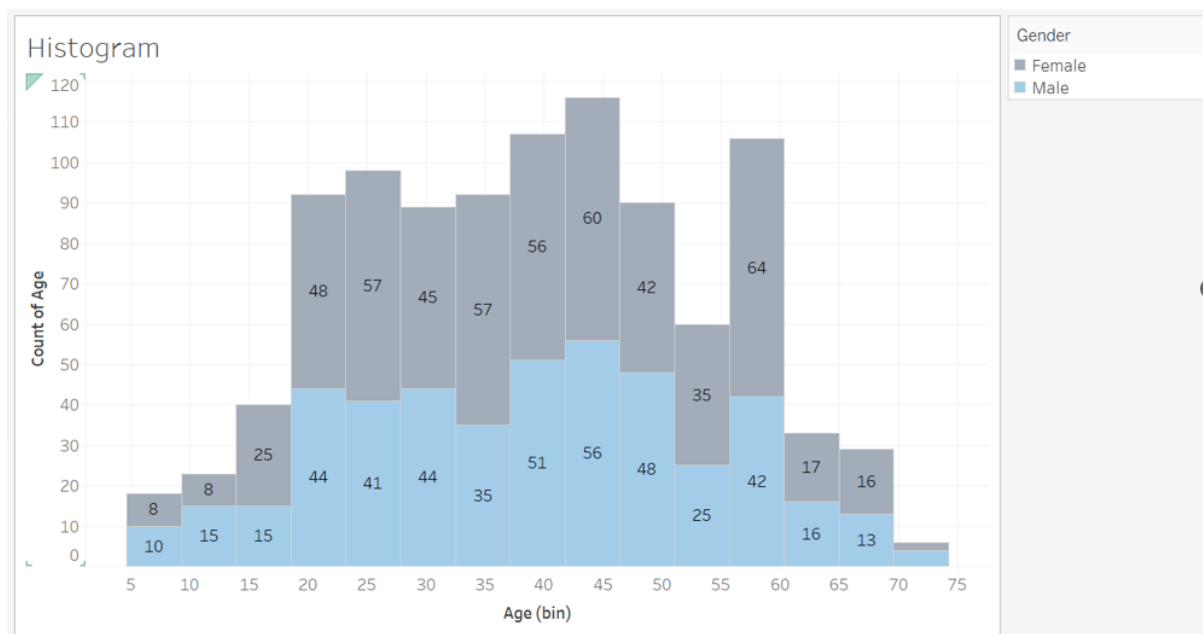
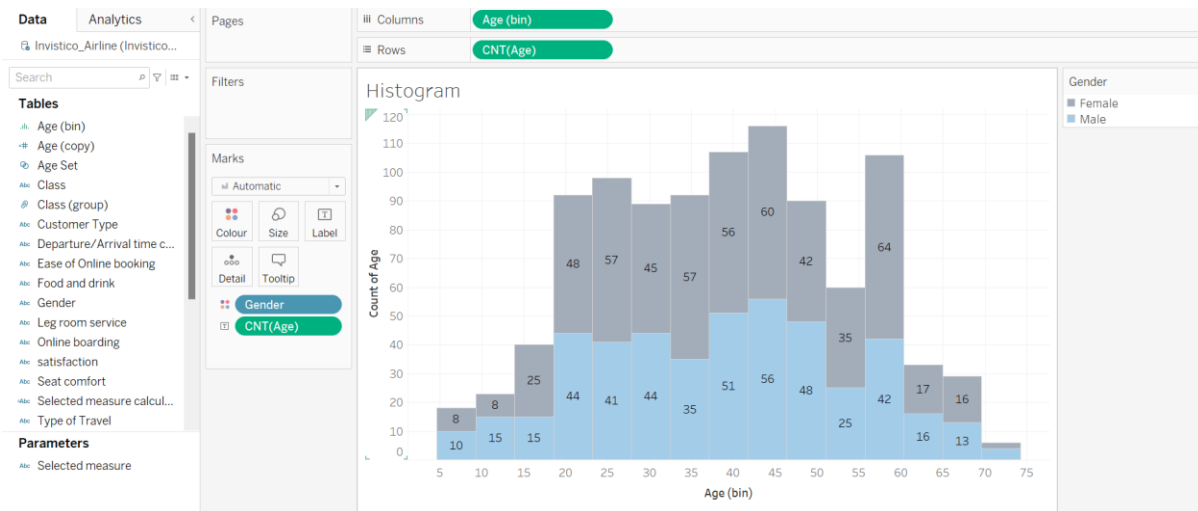
Packed bubbles



Description of the graph

There are 3 classes to compare: Business, Eco and Eco Plus. Every class is represented by different colour (business – purple, eco – orange and eco plus- green). The Eco class has the highest numbers of rows, Eco Plus has the smallest one.

10. Distribution of Age for Men and Women.

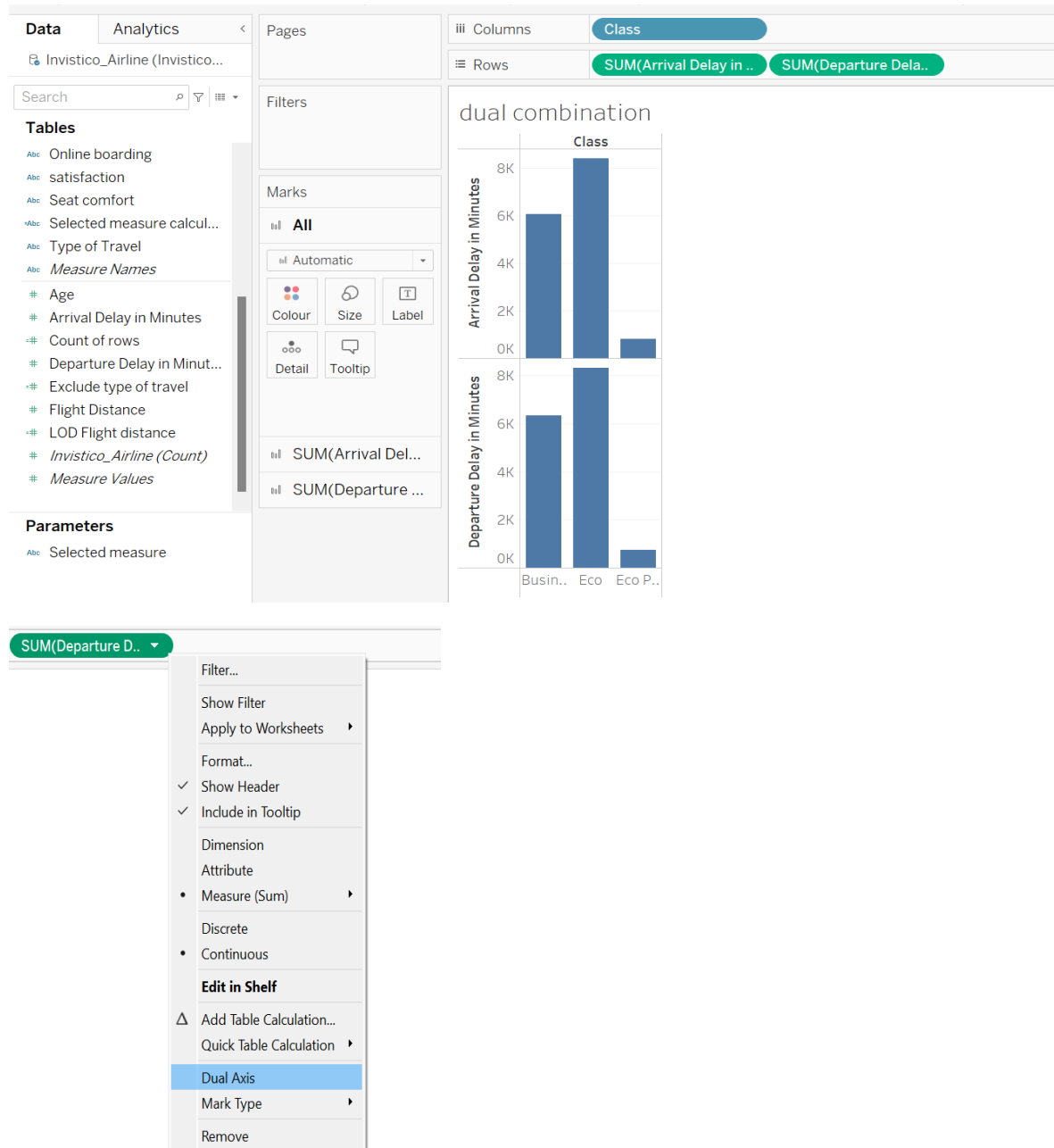


Description of the graph

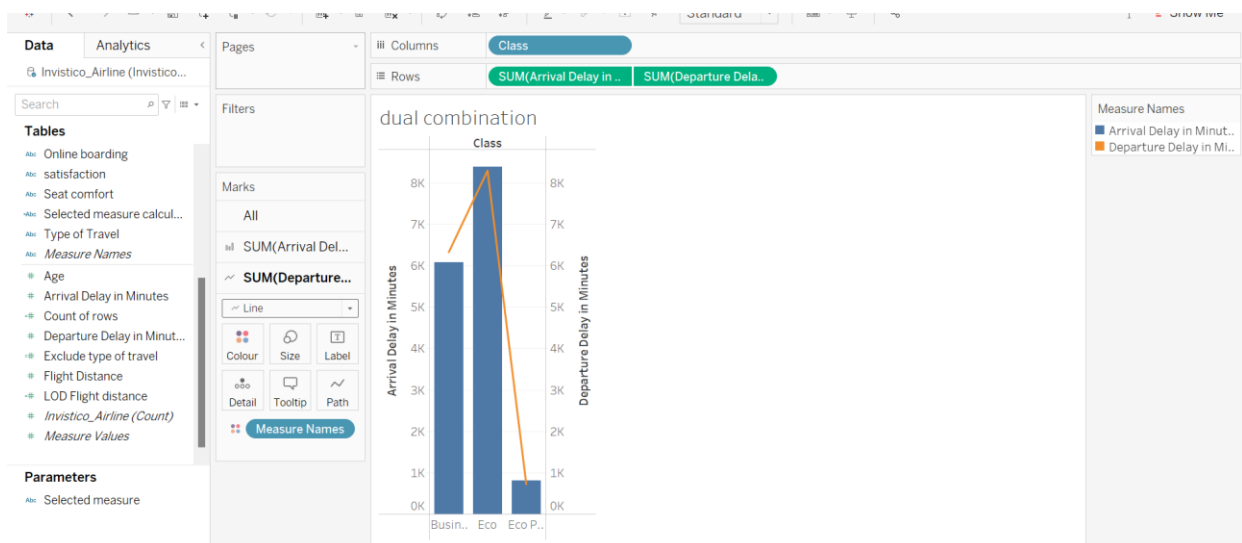
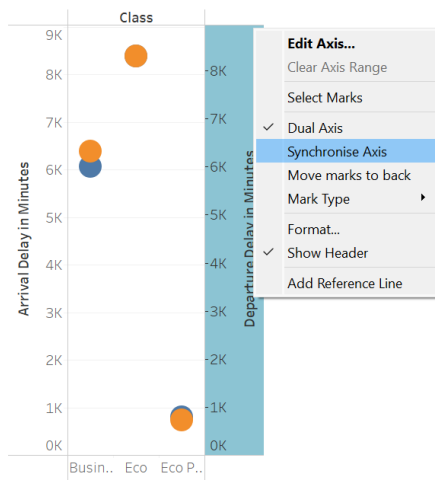
This histogram is made up of row of bars. Each bar has its length and width. The bar's width is a numerical interval which includes a part of observation. The bar's length is an observation strength, in this compartment there are values characteristic to observation strength. On the x-axis there is age (bin), on they-axis there is count of age. The graph shows 15 bins. In each bin it is shown the age of man and woman respectively. The biggest difference between aged man and woman appears in bins: 25 and 65. The highest number of women appear in bin 60 (64) while man in bin 45.

11. Arrival delay in minutes and arrival delay in minutes -dual combination.

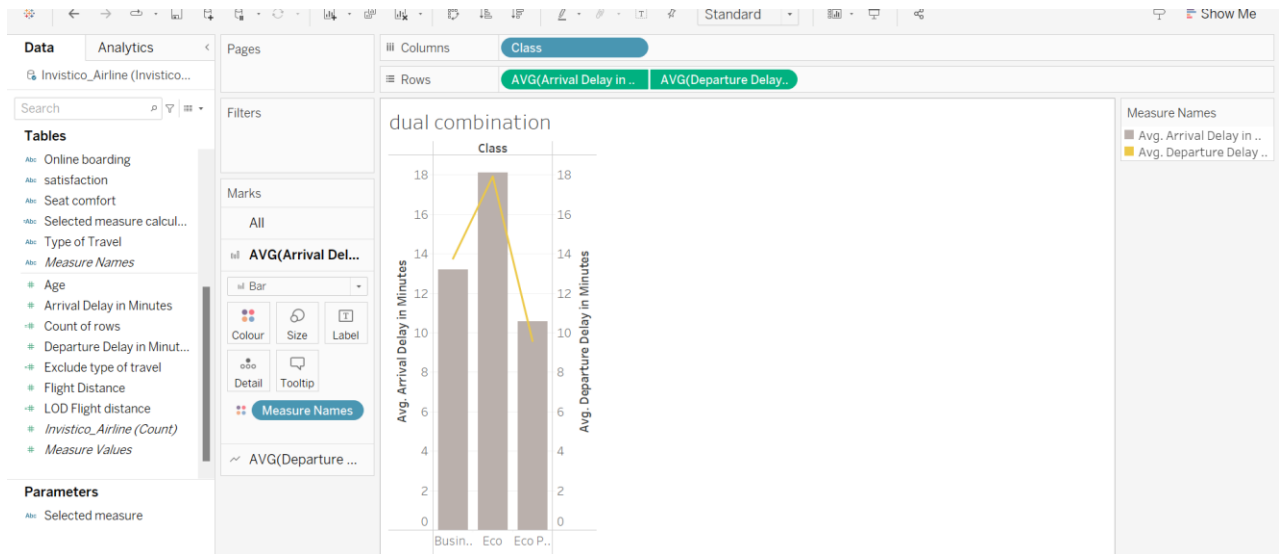
This graph joins 2 various kinds of charts into a single visualization. It allows to collate 2 sets of data which have various scales of measurement. A dual combination chart encompasses a merger of a bar chart and a line chart.



dual combination



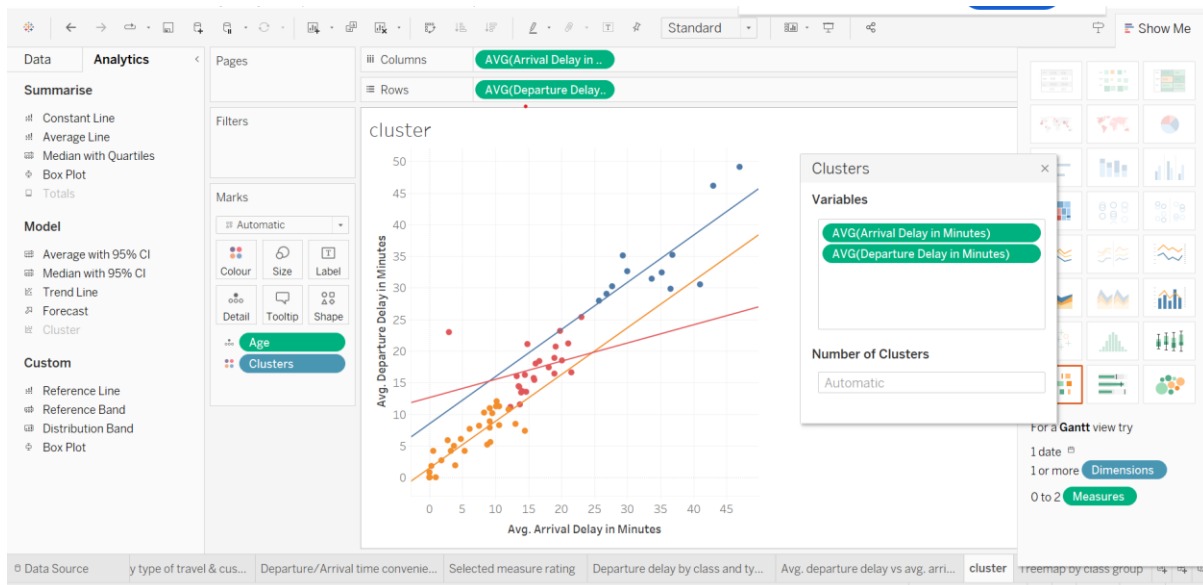
Final graph



Description of the graph

The illustration presents the comparison the average arrival delay in minutes and the average departure delay in minutes. I compare bar versus line; in business class the average arrival delay is smaller than the average departure delay (13207,13759). In Eco class, the average arrival delay is higher than the average departure delay (18117,17924). In Eco plus, the average arrival delay is higher than the average delay (10566, 9579).

12. Cluster analysis of flight delays.





Description of the graph

In this case, Tableau uses algorithm k- means. K-means assigned the observations to every clusters. There are 2 dependencies: the average departure delay in min. and the average arrival delay in min. There are 3 clusters shown on the graph on blue, orange, and red. There is a trend line also.

Data Mining – WEKA - Simple k-means

Choice

I have chosen this data algorithm because it is projected for numeric data (I have numeric and non-numeric values; I have deleted what was not needed for analysis).

How is it working?

There are 5 stages: initialisation, assignment, update, reassignment and update and finalization.

The aim of k-means is to reduce within cluster sum of squared distances betwixt data points and their distinctive cluster centroids. It insistently clarifies the clusters until convergence, where next iterations do not significantly fix the cluster / centroids. This algorithm is sensitive to the primary placing of cluster centroids.

How the process looks like? - more detailed.

- Input details – I have implemented my data set to Weka. I have chosen from cluster option: k-means algorithm. I have only numerical data in my set, as this algorithm doesn't work for non-numerical ones.
- Initial centres- k-means randomly choosing k initial cluster centres from my data set.
- Assign data points- every datapoint in dataset is assigned to the closest cluster centre based on distance.
- Update centres – after assigning data points to clusters, the algorithm, calculate the centre of every cluster built on the mean of the data points belong to cluster.
- Repeating -assigning data points and updating centres until convergence.
- Output clusters – cluster centres (final) are showing centroids of the cluster; every data point is acquainted with corresponding cluster.
- Evaluation- by the usage of different metrics.
- Visualisation – visualise the clusters and centroids.

Anomalies

In this algorithm some anomalies can appear; for example: sensitivity of initialisation, choosing the optimal number of clusters, unequal cluster size, presence of outliers, scaling and normalisation.

In a chosen dataset there are 7 attributes for considerations (preprocess). There are mention below. All the values are numeric.

No.	Name
1	<input type="checkbox"/> Inflight wifi service
2	<input type="checkbox"/> Inflight entertainment
3	<input type="checkbox"/> Online support
4	<input type="checkbox"/> Baggage handling
5	<input type="checkbox"/> Checkin service
6	<input type="checkbox"/> Online boarding
7	<input type="checkbox"/> Departure Delay in Minutes

Percentage split sets on 66% as it gives the best results. The standard option appears the best choice.

Cluster mode

☐ Use training set
☐ Supplied test set
☒ Percentage split %
☐ Classes to clusters evaluation
(Num) Departure Delay in Minutes
☒ Store clusters for visualization

Result list (right-click for options)

11:26:21 - SimpleKMeans
11:29:50 - SimpleKMeans

The results

```

Clusterer output
=== Run information ===

Scheme:      weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -
Relation:    Invistico_Airline - 1k rows (3)-weka.filters.unsupervised.attribute.Remove-R1-6,10-11-weka.filters.unsupervis
Instances:    999
Attributes:   7
              Inflight wifi service
              Inflight entertainment
              Online support
              Baggage handling
              Checkin service
              Online boarding
              Departure Delay in Minutes
Test mode:   split 66% train, remainder test

=== Clustering model (full training set) ===

```

Clusterer output

```

Number of iterations: 8
Within cluster sum of squared errors: 337.4144809233189

Initial starting points (random):

Cluster 0: 4,5,5,4,5,3,0
Cluster 1: 5,2,5,5,4,5,14

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute                                Cluster#
Full Data                                0          1
(999.0)    (576.0)    (423.0)
=====
Inflight wifi service                    3.2282    3.9149    2.2931
Inflight entertainment                  3.3784    4.0503    2.4634
Online support                          3.5045    4.408     2.2742
Baggage handling                        3.7768    3.9497    3.5414
Checkin service                         3.3774    3.783     2.8251
Online boarding                         3.3323    4.1146    2.2671
Departure Delay in Minutes              15.3714   12.7656   18.9196

```

kMeans

=====

```

Number of iterations: 5
Within cluster sum of squared errors: 252.97291254628288

Initial starting points (random):

Cluster 0: 4,4,4,3,2,4,0
Cluster 1: 3,0,3,5,3,3,96

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute                                Cluster#
Full Data                                0          1
(659.0)    (399.0)    (260.0)
=====
Inflight wifi service                    3.2716    3.99     2.1692
Inflight entertainment                  3.3581    3.9799    2.4038
Online support                          3.5023    4.3183     2.25
Baggage handling                        3.7451    3.9023    3.5038
Checkin service                         3.3961    3.7068    2.9192
Online boarding                         3.3642    4.1604    2.1423
Departure Delay in Minutes              15.824   12.9674   20.2077

```

Departure delay in minutes

I am looking for value 20.2077 in cluster 1 when I would like to see the better time delay in this scenario.

Confusion matrix

Skeleton – general

	PREDICTED NO	PREDICTED YES
Actual NO	True Negative	False Positive
Actual YES	False Negative	True Positive

Confusion matrix – interpretation

```
Time taken to build model (percentage split) : 0 seconds
```

```
Clustered Instances
```

```
0      197 ( 58%)  
1      143 ( 42%)
```

197 – These are the cases where the model correctly predicts the negative class

143 – These are the cases where the model incorrectly predicts the negative class when it's actually positive. (Type II error)

58% - These are the cases where the model incorrectly predicts the positive class when it's actually negative. (Type I error)

42% - These are the cases where the model correctly predicts the positive class

To summarise, the model correctly predicts the positive class in 42%.

Interpretation

- Number of iterations

At the beginning, there were 8 iterations; k-means algorithm crosses 8 times to get solution. 8 times were reallocating data points to clusters and updating the cluster centroids. At the end, there were 5 iterations to get final solution.

- Sum of squared errors within cluster

It describes compactness of clusters. 337.4144809233189 means the clusters are not so tight. The smallest values; high tightness.

- Clusters

There are 2 clusters: 0 and 1.

```
Initial starting points (random):
```

```
Cluster 0: 4,5,5,4,5,3,0
```

```
Cluster 1: 5,2,5,5,4,5,14
```

```
Cluster 0: 4,4,4,3,2,4,0
```

```
Cluster 1: 3,0,3,5,3,3,96
```

- Summary

The dataset was divided into a training set and a test set. The k-means model was evaluated on the test split also. The clustering brought about Cluster 0 has 58% of the instances and Cluster 1 has 42% of the instances. Cluster 0 includes 197 instances, which presents circa 58% of the total instances in the dataset. Cluster 1 includes 143 instances, which presents 42% of the total instances in the dataset. Cluster 0 has a higher number of instances than Cluster 1.

Data Ethics

Ethical considerations relating to data analysis

- Safety – analysis should classify passengers', crew, and aircraft safety. Data analysis method should not compromise safety standards.
- Privacy – data anonymization and encryption should be implemented. GDPR and HIPAA should be respected.
- Transparency – stakeholder should understand data sources, methodologies used in analysis.
- Transparency creates accountability.
- Regulatory Compliance- compliance with regulatory command is crucial in aviation data analysis. The analysts must be aware of FAA and GDPR.

Legal considerations relating to data analysis

- Regulatory Compliance – comply with FAA, EASA, GDPR and HIPAA.
- Data Protection Laws – aviation data include personal information about passengers, aircraft, and crew. Data analysis must comply with data protection laws. These ones demand obtaining informed consent from individuals before gathering and using their data.
- Liability – analysing aviation data touch safety and decisions. Organisations need to manage risks. They should notice possible problems, minimize risks, and have insurance to cover any issues.
- Competition and antitrust laws – when aviation competitors divide data for analysis, they are obliged to follow competition laws. They need to avoid actions which could not fair control prices.
- Data security- in aviation data analysis, following data security laws is necessary. Organisations must protect data from unapproved access and cyber threats. They can implement this by encryption, access control and regular checks.

Professional considerations related to data analysis

Professional considerations in aviation data analysis are important for accuracy, ethical conduct, and reliability. Main aspects include possessing expertise and training

in data analysis, assuring accuracy and reliability of analysis results, following to ethical principles for example: integrity, privacy protection adhering to the standards.

Conclusions

TABLEAU

Tableau is a data visualization tool which helps me to produce many visualisations and dashboard from chosen dataset.

I have presented some visualisations in Tableau. I have started from the simplest one (bar chart, pie chart, tree map, scatter plot to the more advanced ones; combined chart, cluster analysis and LoD calculations).

The “Pie Chart by Gender” as one of the simplest ones presents the percentage of men and woman in the entire dataset. Female share is slightly higher than male (8.1% difference).

The “Departure Delay by Class and Type of Travel” displays average departure delay in minutes for different classes (Business, Eco, Eco Plus) and types of travel (Business, Personal).

The “Flight Distance by Customer and Gender” compares flight distance by customer type and gender, divided into disloyal and loyal customers. Utilizes a reference line for comparison.

The “Tree Map by Class Group” exhibits hierarchical data of Eco, Eco Plus, and Business classes. Rectangles represent satisfaction levels (satisfied or dissatisfied).

The “Average Departure Delay vs. Arrival Delay by Age” illustrates the dependency between average departure delay and average arrival delay by age, with a strong linear trend line.

The “Average Flight Distance vs. Average Delay by Age” depicts the relationship between average arrival delay and flight distance, with age as a conditioning factor.

The “Comparison of classes: Business, Eco and Eco Plus” contrasts classes using circles' cluster, with each class represented by a different colour.

The “Distribution of Age for Men and Women” highlights the age distribution for men and women using histogram bars.

The “Arrival Delay in Minutes and Arrival Delay in Minutes” collates average arrival delay and departure delay using both bar and line charts.

The “Cluster Analysis of Flight Delays” utilizes k-means algorithm to analyze flight delays, showing 3 clusters and a trend line.

The important findings

- each visualisation serves a specific purpose,
- there are simple and complex charts,
- not all charts in Tableau will be possible to create due to the chosen dataset.

WEKA - Simple k-means algorithm.

Cluster Profiles

Two clusters: 0 and 1, Cluster 0 represents passengers who generally rated their in flight experience higher, with higher ratings for inflight Wi-Fi service, inflight entertainment, online support, baggage handling, check in service and online boarding. Cluster 1 represents passengers who gave lower ratings across these attributes compared to cluster 0.

Attribute Analysis

Passengers in Cluster 0 tend to rate their in-flight experience more positively, indicating higher satisfaction with various services provided by the airline.

Passengers in Cluster 1 gave lower rating across most attributes, suggesting lower satisfaction levels with the airline services.

Departure Delay

Cluster 0 experienced shorter average departure delays compared to Cluster 1. This could highlight that passengers in Cluster 0 had smoother and more punctual flight experiences in terms of departure times.

Model Performance

The clusters' centroids were computed based on the provided attributes and used to assign instances to clusters.

The within cluster sum of squared errors decreased from the full training set to the test split, indicating improved clustering performance on unseen data.

Cluster Distribution

The data is somewhat skewed, with Cluste 0 containing a higher number of instances (58%) compared to Cluster 1 (42%).

The usage of obtained data

The presented data can be used by airlines, travel agencies and aviation analysts.

- Firstly, airline companies can use these findings to better understand customers preferences and satisfaction levels. They can adjust their services and marketing options to entice and retain customers.
- Secondly, travel agencies can utilise these findings to offer individualised travel packages and suggestions to their clientele based on their preferences and behaviour trends.
- Thirdly, aviation analysts can analyse these results to obtain insights into trends and patterns in flight delays, travel behaviour and customer satisfaction.

How these findings can be used?

- Marketing Strategies – adjust marketing options based on customer preferences. For instance, they can focus on particular segments with individualised promotions.
- Customer Retention – airlines can boost customer satisfaction by addressing areas of concern.
- Service Enhancement – airlines and airport authorities can use these data to recognise fields for service improvement (for example streamline check-in process, improving overall customer experience)

