NYCFLIGHTS REPORT

Report Overview

This report was created for an overview quality diagnosis of **.** data. It was created for **the purpose of judging the validity of variables** before conducting EDA.

Contents

Overview

Data Structures

division	metrics	value
size	observations	336,776
size	variables	19
size	values	6,398,744
size	memory size (MB)	39
duplicated	duplicate observation	0
missing	complete observation	327,346
missing	missing observation	9,430
missing	missing variables	6
missing	missing values	46,595

division	metrics	value
data type	numerics	6
data type	integers	8
data type	factors/ordered	0
data type	characters	4
data type	Dates	0
data type	POSIXcts	1
data type	others	0

Table 1: Data structures and types

Job Informations

division	metrics	value
dataset	dataset	.
dataset	dataset type	tbl_df
job	samples	336,776 / 336,776 (100%)
job	created	27 Nov, 2022
job	created by	George Ngugi

Table 2: Job informations

Warnings

checks	judgements	removes
5	11	1

Table 3: Summary of warnings

warnings	status	recommand
arr_delay has 9,430 (2.8%) missing values	missing	judgement
air_time has 9,430 (2.8%) missing values	missing	judgement
arr_time has 8,713 (2.6%) missing values	missing	judgement
dep_time has 8,255 (2.5%) missing values	missing	judgement
dep_delay has 8,255 (2.5%) missing values	missing	judgement
tailnum has 2,512 (0.7%) missing values	missing	judgement
year has constant value "2013"	cardinality	remove
minute has 60,696 (18.02%) zeros	zero	check
dep_delay has 16,514 (4.9%) zeros	zero	check
arr_delay has 5,409 (1.61%) zeros	zero	check
arr_delay has 188,933 (56.1%) negatives	negative	check
dep_delay has 183,575 (54.51%) negatives	negative	check
dep_delay has 43,216 (12.83%) outliers	outlier	judgement
arr_delay has 27,880 (8.28%) outliers	outlier	judgement
air_time has 5,448 (1.62%) outliers	outlier	judgement
distance has 715 (0.21%) outliers	outlier	judgement
flight has 1 (0%) outliers	outlier	judgement

Table 4: Warnings in dataset and variables

Variables

variables	types	missing	cardinality	zero	minus	outlier
year	integer		constant			
month	integer					
day	integer					
dep_time	integer	X				
sched_dep_time	integer					
dep_delay	numeric	X		X	X	X
arr_time	integer	X				
sched_arr_time	integer					
arr_delay	numeric	X		X	X	X
carrier	character					
flight	integer					X
tailnum	character	X				
origin	character					
dest	character					
air_time	numeric	X				X
distance	numeric					X
hour	numeric					
minute	numeric			X		
time_hour	POSIXct					

Table 5: List of variables diagnosis

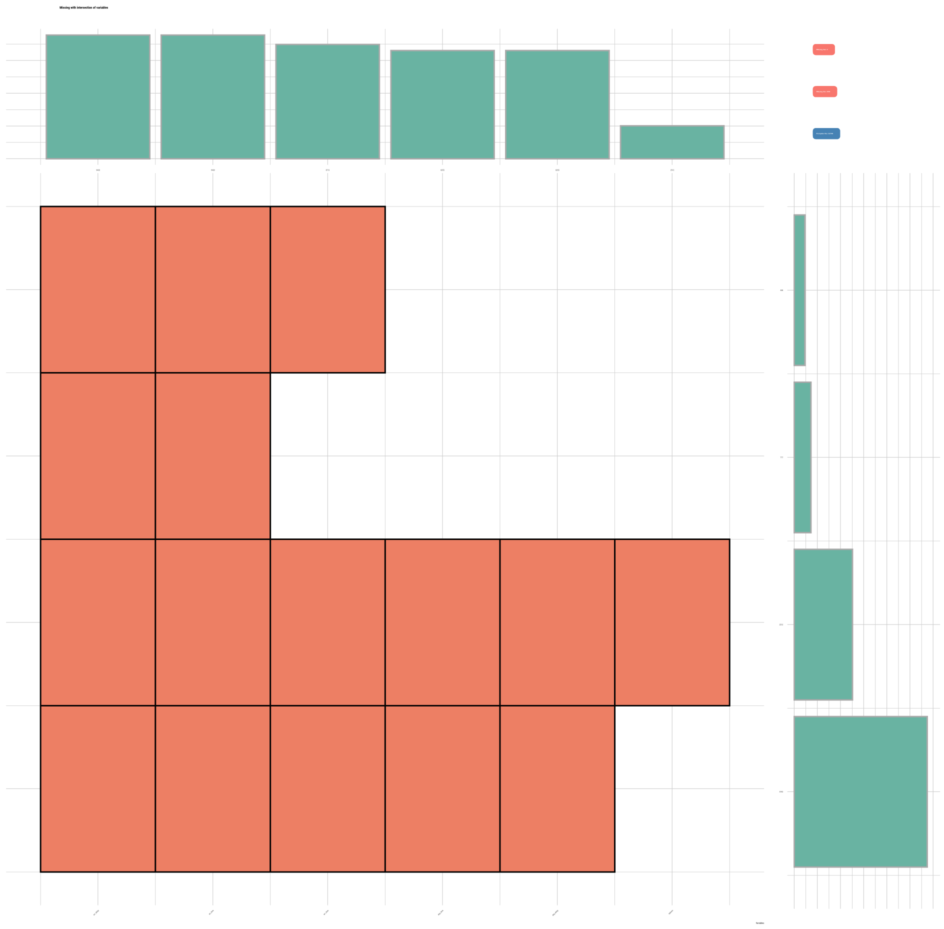
Missing Values

List of Missing Values

variables	missing_count	missing (%)	status	recommand
arr_delay	9,430	2.8%	Good	Delete or Imputation
air_time	9,430	2.8%	Good	Delete or Imputation
arr_time	8,713	2.6%	Good	Delete or Imputation
dep_time	8,255	2.5%	Good	Delete or Imputation
dep_delay	8,255	2.5%	Good	Delete or Imputation
tailnum	2,512	0.7%	Good	Delete or Imputation

Table 6: List of variables including missing values

Visualization



Unique Values

Categorical Vaiables

No variable with a high proportion greater than 0.5

Numerical Vaiables

Variables where the unique cases is less than 5 or unique is 1.

variables	types	unique	unique (%)	status	recommand
year	integer	1	0%	constant	Remove Variable

Table 7: Detail warning numerical cardinality

Categorical Variable Diagnosis

Top Ranks

variables	levels	freq	ratio (%)
carrier	UA	58,665	17.4
carrier	B6	54,635	16.2
carrier	EV	54,173	16.1
carrier	DL	48,110	14.3
carrier	AA	32,729	9.7
carrier	MQ	26,397	7.8
carrier	US	20,536	6.1
carrier	9E	18,460	5.5
carrier	WN	12,275	3.6
carrier	VX	5,162	1.5
carrier	Other levles	5,634	1.7
dest	ORD	17,283	5.1
dest	ATL	17,215	5.1
dest	LAX	16,174	4.8
dest	BOS	15,508	4.6
dest	MCO	14,082	4.2
dest	CLT	14,064	4.2
dest	SFO	13,331	4.0
dest	FLL	12,055	3.6
dest	MIA	11,728	3.5
dest	DCA	9,705	2.9
dest	Other levles	195,631	58.1
origin	EWR	120,835	35.9
origin	JFK	111,279	33.0
origin	LGA	104,662	31.1

Table 8: Top 10 levels of categorical variables

variables	levels	freq	ratio (%)
tailnum	N725MQ	575	0.2
tailnum	N722MQ	513	0.2
tailnum	N723MQ	507	0.2
tailnum	N711MQ	486	0.1
tailnum	N713MQ	483	0.1
tailnum	N258JB	427	0.1
tailnum	N298JB	407	0.1
tailnum	N353JB	404	0.1
tailnum	N351JB	402	0.1
tailnum	Other levles	330,060	98.0
tailnum	Missing	2,512	0.7
time_hour	2013-09-13 08:00:00	94	0.0
time_hour	2013-09-20 08:00:00	94	0.0
time_hour	2013-09-09 08:00:00	93	0.0
time_hour	2013-09-16 08:00:00	93	0.0
time_hour	2013-09-23 08:00:00	93	0.0
time_hour	2013-09-19 08:00:00	92	0.0
time_hour	2013-10-11 08:00:00	92	0.0
time_hour	2013-09-10 08:00:00	91	0.0
time_hour	2013-09-12 08:00:00	91	0.0
time_hour	2013-09-17 08:00:00	91	0.0
time_hour	Other levles	335,852	99.7

Table 8: Top 10 levels of categorical variables (continued)

Numerical Variable Diagnosis

Distributions

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
year	2,013	2,013	2,013.00	2,013	2,013	2,013	0	0	0
month	1	4	6.55	7	10	12	0	0	0
day	1	8	15.71	16	23	31	0	0	0
dep_time	1	907	1,349.11	1,401	1,744	2,400	0	0	0
sched_dep_time	106	906	1,344.25	1,359	1,729	2,359	0	0	0
dep_delay	-43	-5	12.64	-2	11	1,301	16,514	183,575	43,216
arr_time	1	1,104	1,502.05	1,535	1,940	2,400	0	0	0
sched_arr_time	1	1,124	1,536.38	1,556	1,945	2,359	0	0	0
arr_delay	-86	-17	6.90	-5	14	1,272	5,409	188,933	27,880
flight	1	553	1,971.92	1,496	3,465	8,500	0	0	1
air_time	20	82	150.69	129	192	695	0	0	5,448
distance	17	502	1,039.91	872	1,389	4,983	0	0	715
hour	1	9	13.18	13	17	23	0	0	0
minute	0	8	26.23	29	44	59	60,696	0	0

Table 9: General list of numerical diagnosis

Zero Values

variables	min	median	max	zero	zero (%)
minute	0	29	59	60,696	18.0
dep_delay	-43	-2	1,301	16,514	4.9
arr_delay	-86	-5	1,272	5,409	1.6

Table 10: List of numerical diagnosis (zero)

Negative Values

variables	min	median	max	minus	minus (%)
arr_delay	-86	-5	1,272	188,933	56.1
dep_delay	-43	-2	1,301	183,575	54.5

Table 11: List of numerical diagnosis (minus)

Outliers

List of Outliers

variables	min	median	max	outlier	outlier (%)
dep_delay	-43	-2	1,301	43,216	12.8
arr_delay	-86	-5	1,272	27,880	8.3
air_time	20	129	695	5,448	1.6
distance	17	872	4,983	715	0.2
flight	1	1,496	8,500	1	0.0

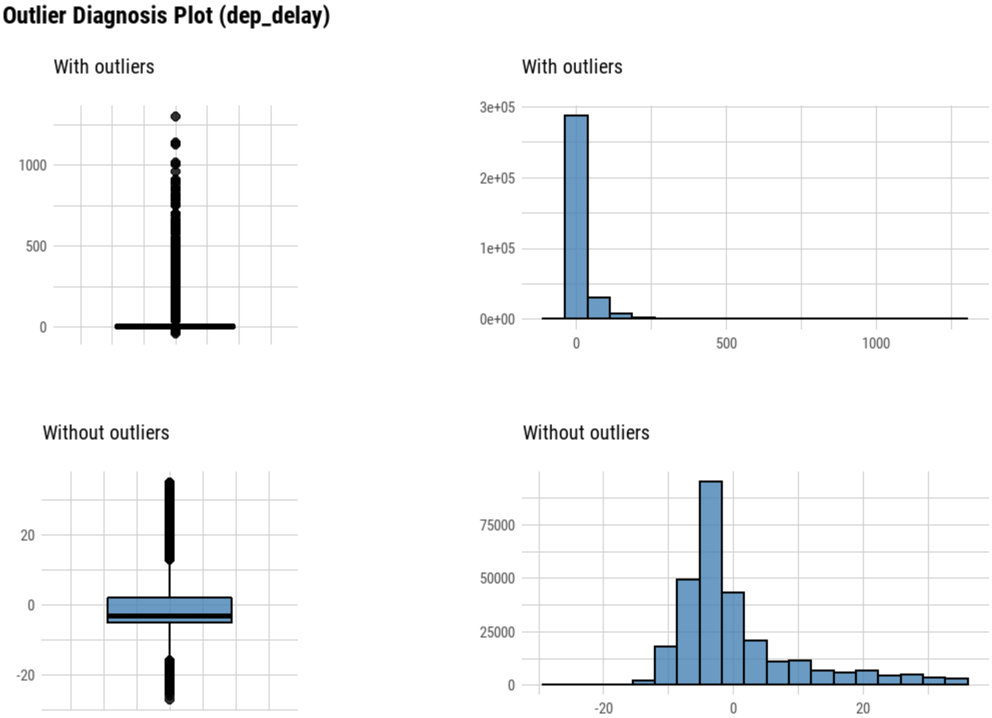
Table 12: Diagnosis of numerical variable outliers

Individual Outliers

variable: dep_delay

Measures	Values
Outliers count	43,216
Outliers ratio (%)	12.83%
Mean of outliers	93.14666
Mean with outliers	12.63907
Mean without outliers	0.4443455

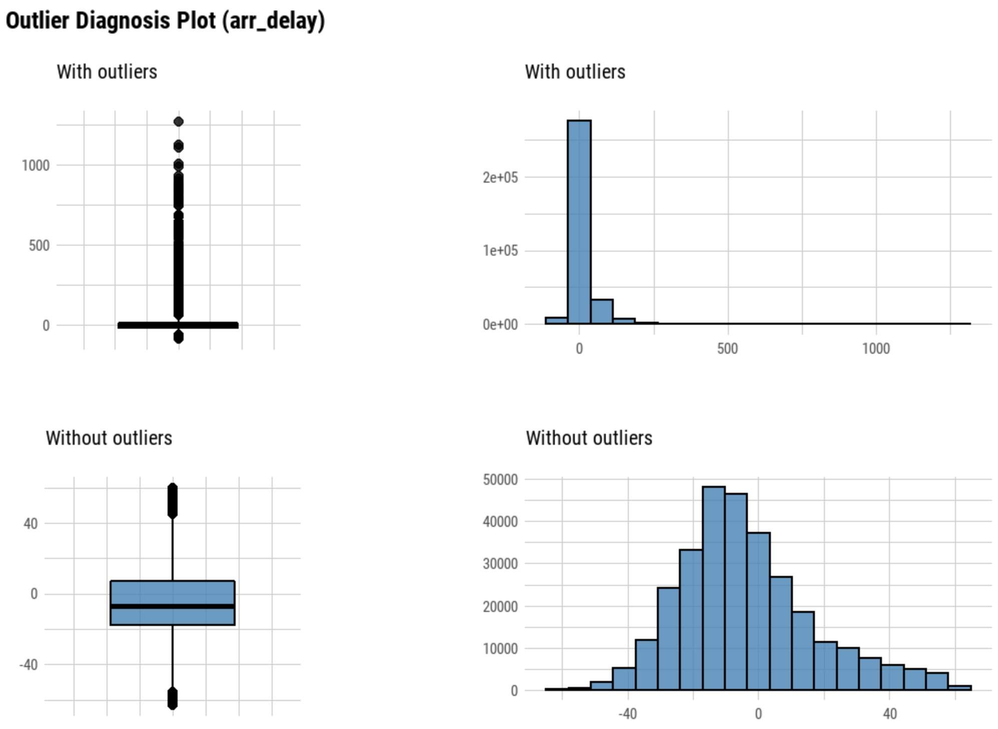
Table 13: dep_delay



variable: arr_delay

Measures	Values
Outliers count	27,880
Outliers ratio (%)	8.28%
Mean of outliers	120.5562
Mean with outliers	6.895377
Mean without outliers	-3.686342

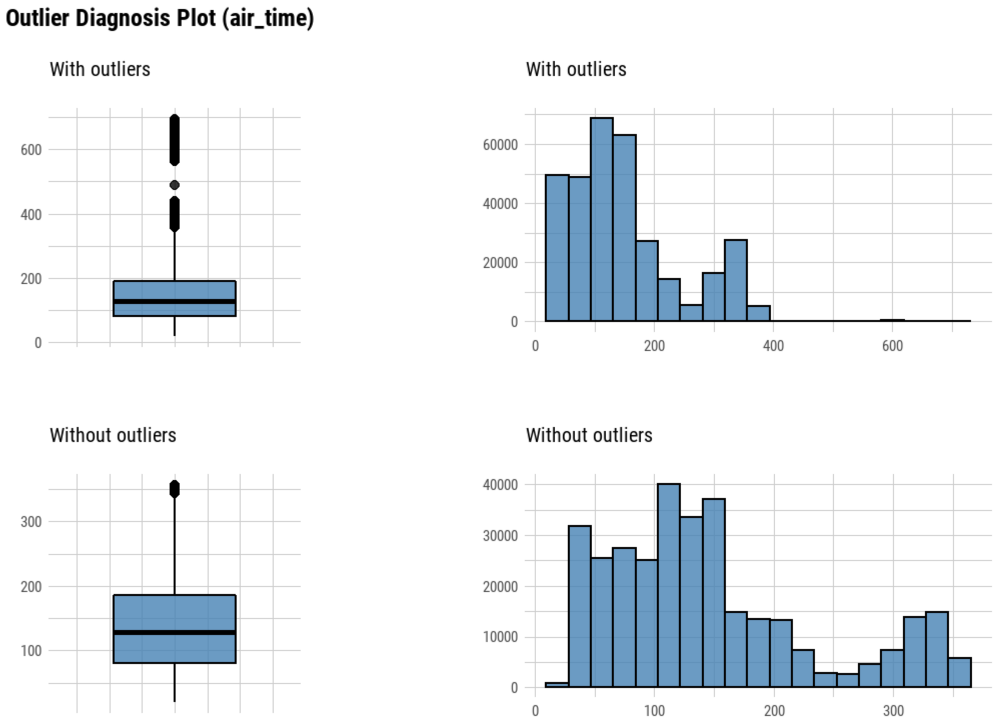
Table 13: arr_delay



variable: air_time

Measures	Values
Outliers count	5,448
Outliers ratio (%)	1.62%
Mean of outliers	400.1419
Mean with outliers	150.6865
Mean without outliers	146.4645

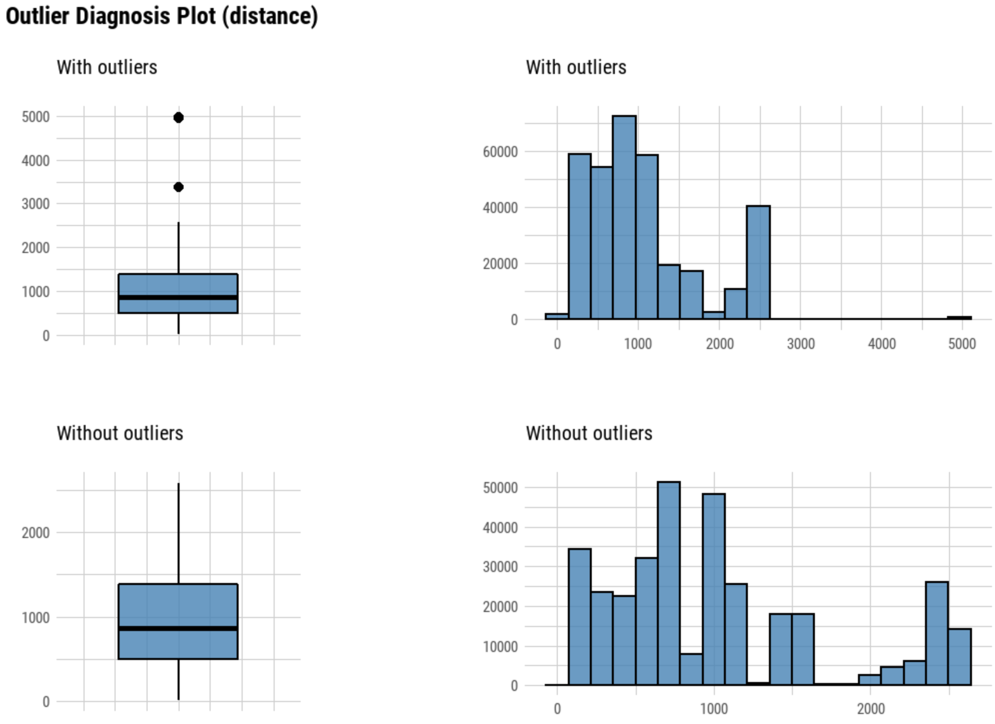
Table 13: air_time



variable: distance

Measures	Values
Outliers count	715
Outliers ratio (%)	0.21%
Mean of outliers	4954.743
Mean with outliers	1039.913
Mean without outliers	1031.583

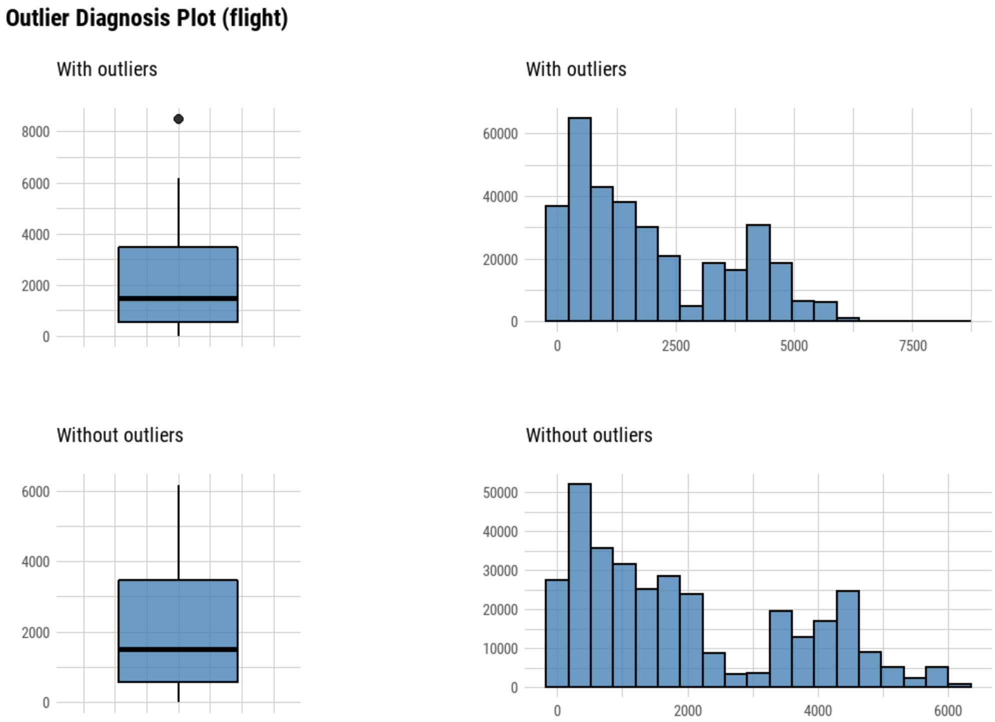
Table 13: distance



variable: flight

Measures	Values
Outliers count	1
Outliers ratio (%)	0%
Mean of outliers	8500
Mean with outliers	1971.924
Mean without outliers	1971.904

Table 13: flight


Outlier Diagnosis Plot (flight)
With outliers
With outliers
Without outliers
Without outliers