EDA REPORT

Report Overview

This report was created for the EDA of . data. It helps explore data to **understand the data and find scenarios for performing the analysis.**

Contents

Overview

Data Structures

division	metrics	value
size	observations	4,925
size	variables	7
size	values	34,475
size	memory size (KB)	0
duplicated	duplicate observation	412
missing	complete observation	4,925
missing	missing observation	0
missing	missing variables	0
missing	missing values	0

division	metrics	value
data type	numerics	3
data type	integers	0
data type	factors/ordered	0
data type	characters	2
data type	Dates	2
data type	POSIXcts	0
data type	others	0

Table 1: Data structures and types

Job Informations

division	metrics	value
dataset	dataset	.
dataset	dataset type	spec_tbl_df
dataset	target	not defied
job	samples	4,925 / 4,925 (100%)
job	created	2022-02-17 11:03:50
job	created by	dlookr

Table 2: Job informations

Univariate Analysis

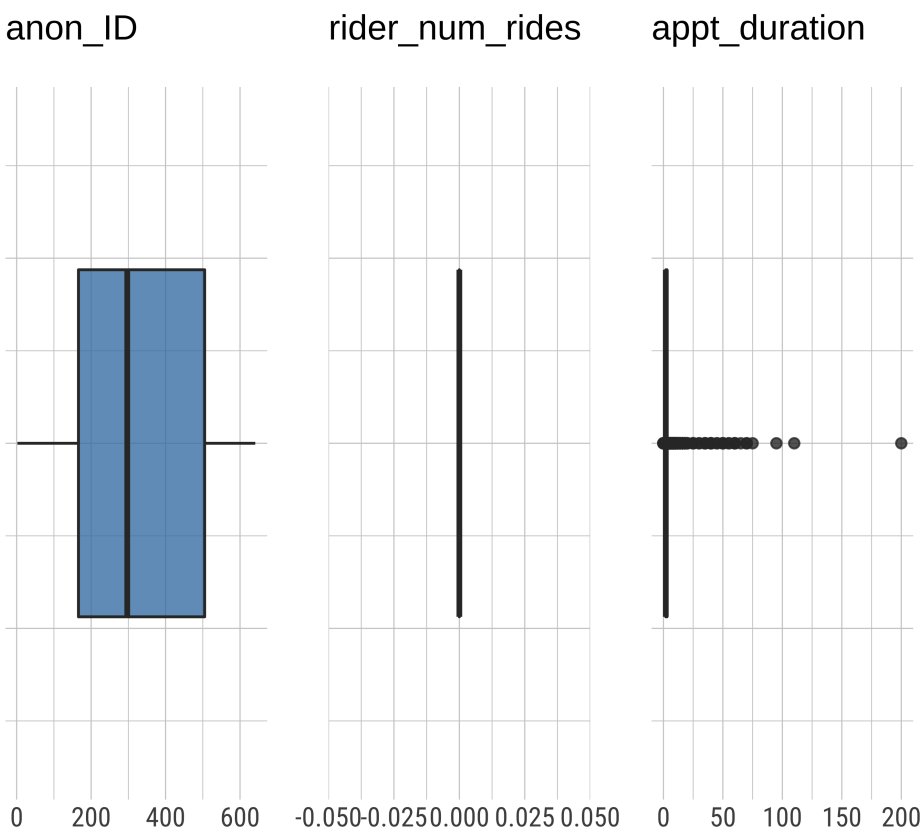
Descriptive Statistics

Numerical Variables

variables	missing	mean	sd	min	Q1	median	Q3	max
anon_ID	0	329.86	187.98	2	166	297	505	641
rider_num_rides	0	0.00	0.00	0	0	0	0	0
appt_duration	0	3.17	5.65	0	2	2	3	200

Table 3: Descriptive statistics of numerical variables

Distribution by numerical variables



variables	data types	distinct	skewness	kurtosis	zero	negative	outlier
anon_ID	numeric	162	0.15	-1.24	0	0	0
rider_num_rides	numeric	1	NaN	NaN	4,925	0	0
appt_duration	numeric	56	15.20	377.29	25	0	650

Categorical Variables

variables	levels	observations	frequency	frequency(%)	rank
rider_first_ride_date	2015-05-06	4,925	548	11.13	1
rider_first_ride_date	2015-05-02	4,925	326	6.62	2
rider_first_ride_date	2015-04-20	4,925	295	5.99	3
rider_first_ride_date	2015-04-27	4,925	226	4.59	4
rider_first_ride_date	2015-05-04	4,925	154	3.13	5
rider_first_ride_date	2015-04-23	4,925	120	2.44	6
rider_first_ride_date	2015-04-21	4,925	118	2.40	7
rider_first_ride_date	2015-05-08	4,925	111	2.25	8
rider_first_ride_date	2016-01-13	4,925	104	2.11	9
rider_first_ride_date	2016-01-15	4,925	102	2.07	10
rider_last_ride_date	2021-08-31	4,925	802	16.28	1
rider_last_ride_date	2021-07-21	4,925	535	10.86	2
rider_last_ride_date	2021-10-18	4,925	187	3.80	3
rider_last_ride_date	2021-12-23	4,925	154	3.13	4
rider_last_ride_date	2021-10-06	4,925	119	2.42	5
rider_last_ride_date	2021-11-04	4,925	118	2.40	6
rider_last_ride_date	2021-11-03	4,925	116	2.36	7
rider_last_ride_date	2021-10-22	4,925	114	2.31	8
rider_last_ride_date	2021-10-15	4,925	102	2.07	9
rider_last_ride_date	2021-11-16	4,925	93	1.89	10
appt_date	1/27/2020	4,925	59	1.20	1
appt_date	1/31/2019	4,925	51	1.04	2
appt_date	9/28/2020	4,925	45	0.91	3
appt_date	11/30/2020	4,925	39	0.79	4
appt_date	10/28/2019	4,925	38	0.77	5

Table 4: Top rank levels of categorical variables

variables	levels	observations	frequency	frequency(%)	rank
variables	levels	observations	frequency	frequency(%)	rank
appt_date	4/30/2021	4,925	36	0.73	6
appt_date	7/30/2021	4,925	36	0.73	6
appt_date	3/31/2021	4,925	35	0.71	8
appt_date	12/30/2020	4,925	33	0.67	9
appt_date	10/29/2019	4,925	31	0.63	10
category	Doctor Appt	4,925	2,102	42.68	1
category	Shopping	4,925	1,158	23.51	2
category	Board or Committee Mtg	4,925	586	11.90	3
category	Friendly Visit	4,925	571	11.59	4
category	Pantry	4,925	273	5.54	5
category	Errands	4,925	79	1.60	6
category	Special Projects	4,925	56	1.14	7
category	Odd Jobs	4,925	43	0.87	8
category	Pantry Delivery	4,925	30	0.61	9
category	Skilled Work	4,925	20	0.41	10

Table 4: Top rank levels of categorical variables (continued)

The number of categorical(factor/ordered) variables is 0.

Normality Test

variable	min	Q1	median	Q3	max	skewness	kurtosis	balance
anon_ID	2	166	297	505	641	0.1	-1.2	Balanced
rider_num_rides	0	0	0	0	0	NaN	NaN	Invalid
appt_duration	0	2	2	3	200	15.2	377.3	Right-Skewed

Table 5: Descriptive statistics of numerical variables

anon_ID

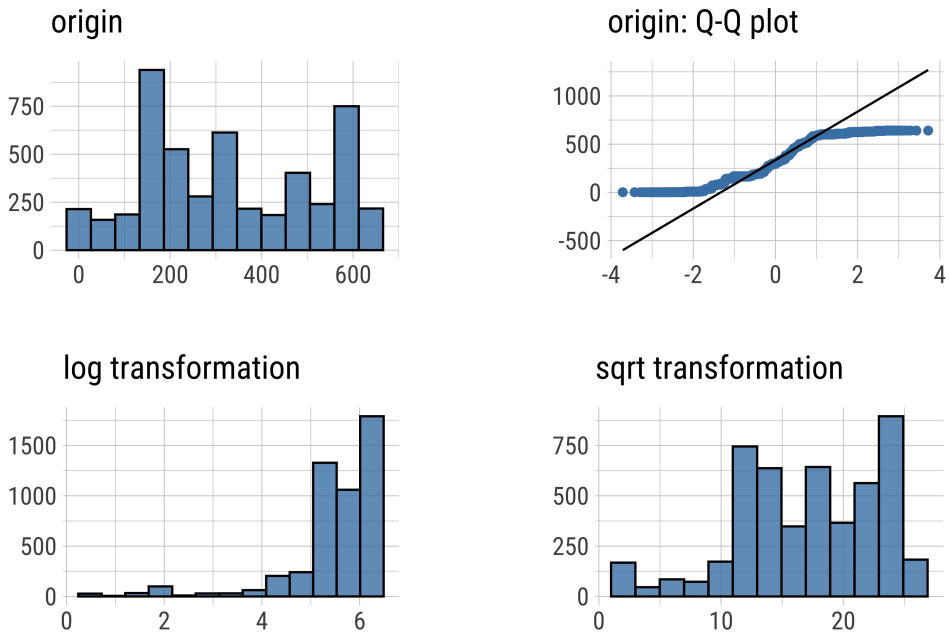
statistic	p_value	remark
0.93165	8.956e-43	No sample

Table 6: Shapiro-Wilk normality test

type	skewness	kurtosis
original	0.1484	1.7630
log transformation	-2.2667	9.4700
sqrt transformation	-0.5209	2.7342

Table 6: skewness and kurtosis

Normality Diagnosis Plot (x)



rider_num_rides

(unique) sample size must be greater then 3

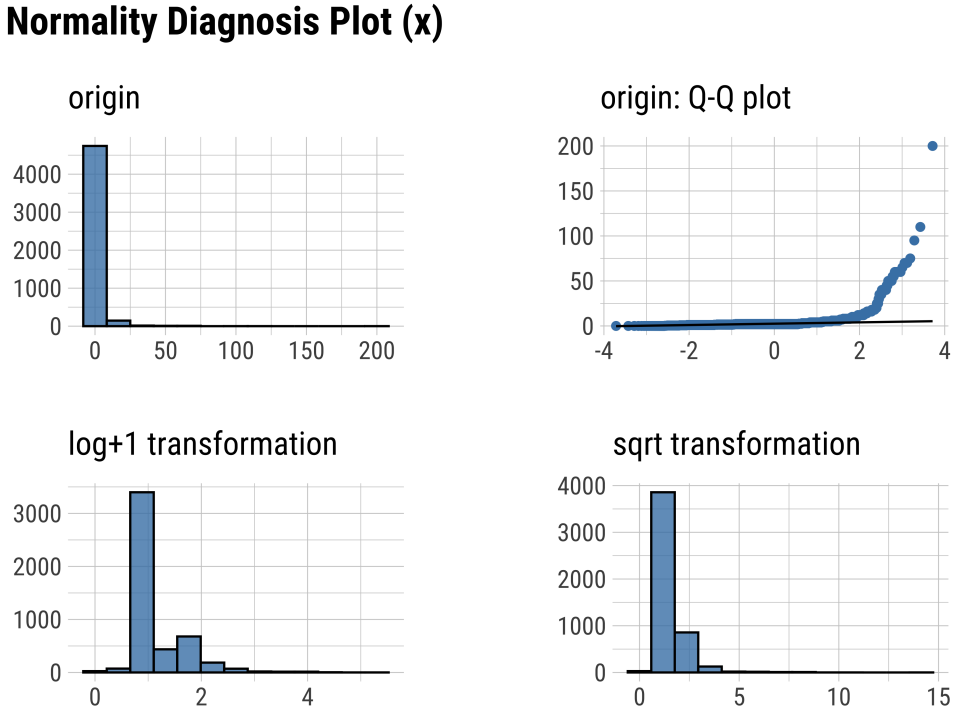
appt_duration

statistic	p_value	remark
0.24374	5.8127e-89	No sample

Table 6: Shapiro-Wilk normality test

type	skewness	kurtosis
original	15.1920	379.9050
log+1 transformation	2.2562	12.1256
sqrt transformation	5.0499	48.5419

Table 6: skewness and kurtosis



Bivariate Analysis

Compare Numerical Variables

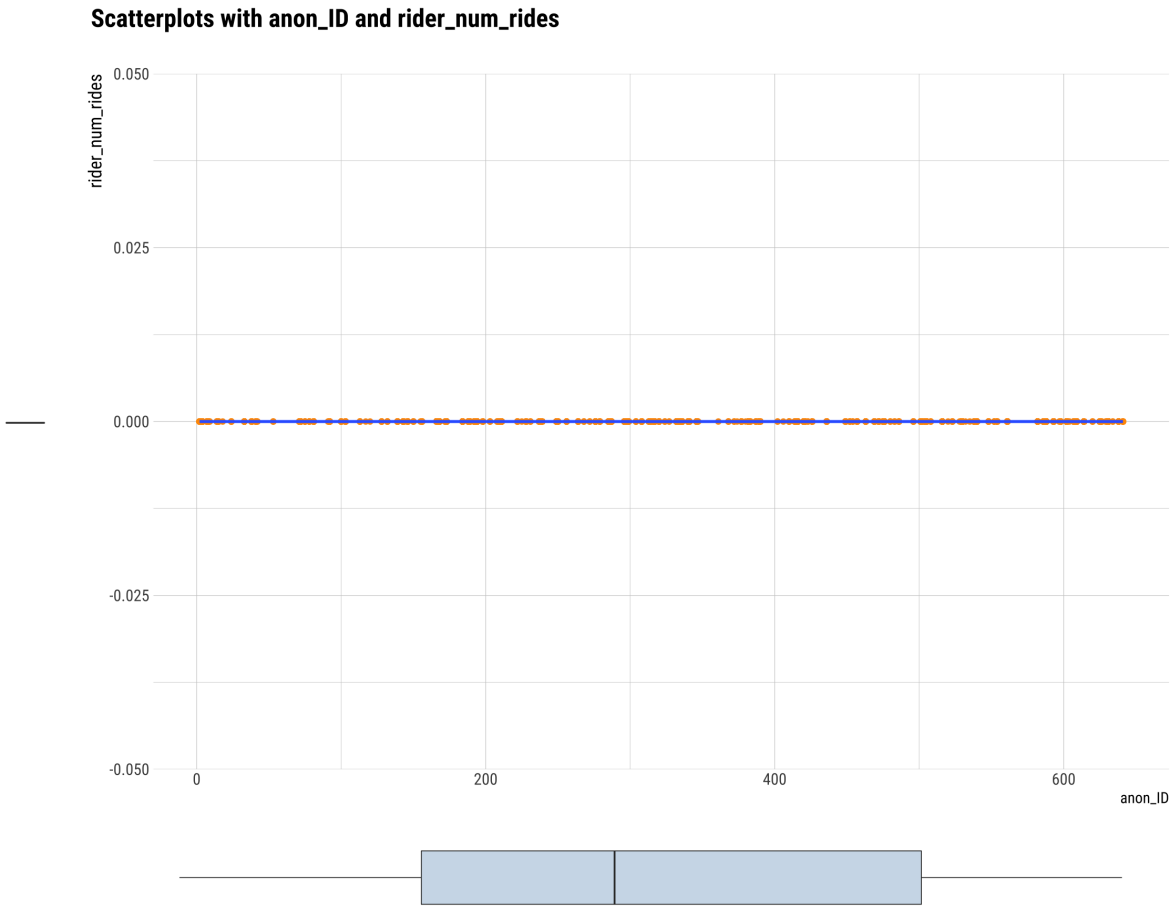
first variable	second variable	correlation coefficient
anon_ID	rider_num_rides	NA
anon_ID	appt_duration	0.15368
rider_num_rides	appt_duration	NA

Table 7: Correlation coefficient

‘anon_ID’ vs ‘rider_num_rides’

first variable	second variable	r.squared	adj.r.squared	sigma	statistic	p.value	df
anon_ID	rider_num_rides	0	0	187.983	NA	NA	NA

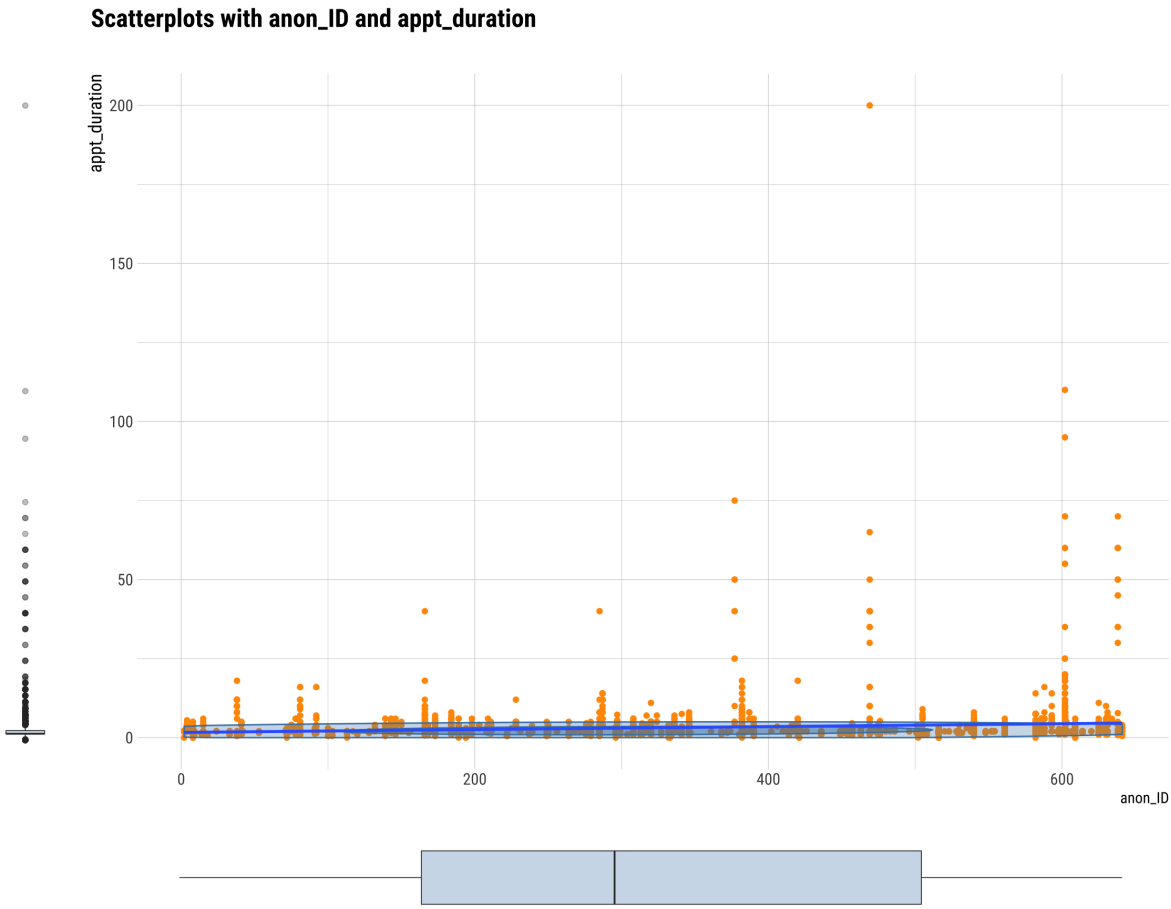
Table 7: Summary of linear model



‘anon_ID’ vs ‘appt_duration’

first variable	second variable	r.squared	adj.r.squared	sigma	statistic	p.value	df
anon_ID	appt_duration	0.0236188	0.0234205	185.7686	119.0882	0	1

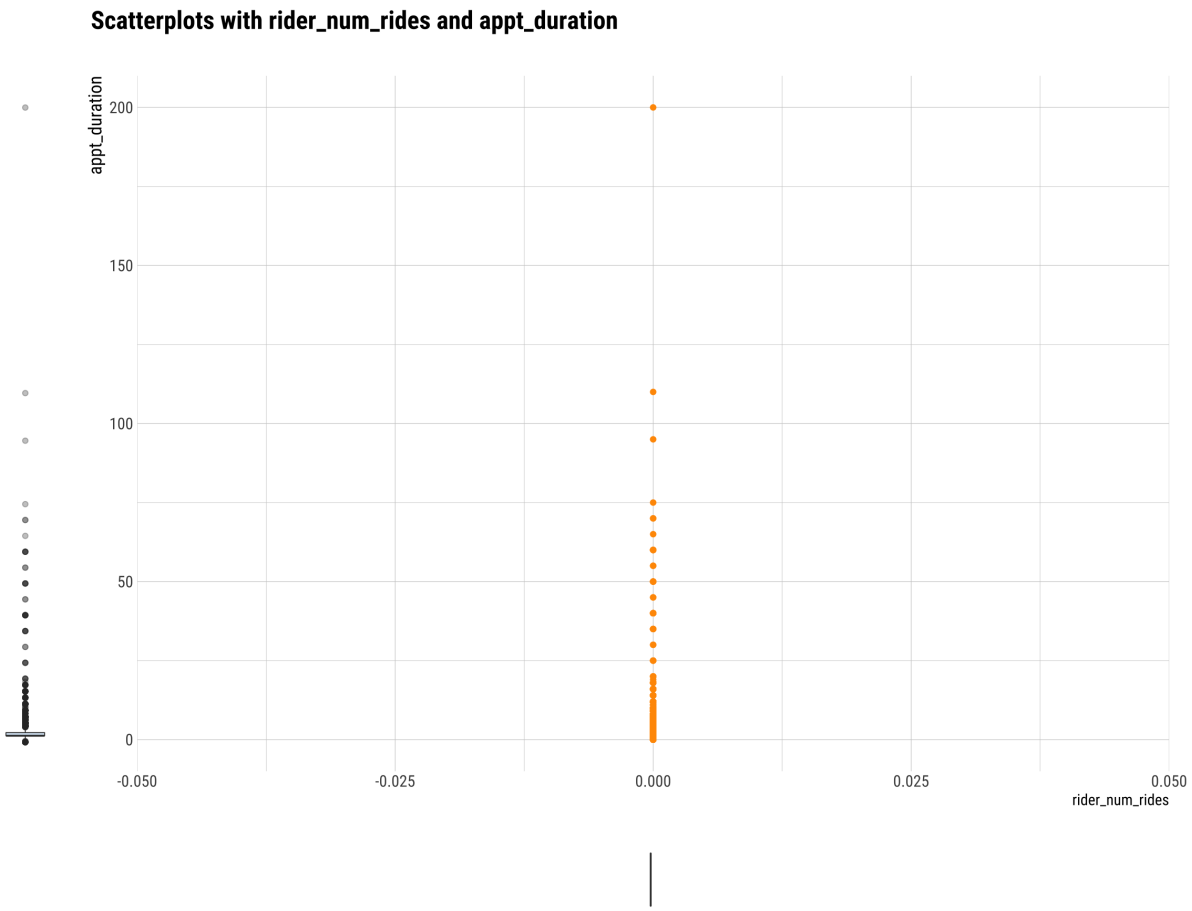
Table 7: Summary of linear model



‘rider_num_rides’ vs ‘appt_duration’

first variable	second variable	r.squared	adj.r.squared	sigma	statistic	p.value	df
rider_num_rides	appt_duration	NaN	NaN	0	NaN	NA	1

Table 7: Summary of linear model



Compare Categorical Variables

The number of categorical variables is less than 2.

Multivariate Analysis

Correlation Analysis

Correlation Coefficient Matrix

	second variable		
first variable	**anon_ID**	**rider_num_rides**	**appt_duration**
anon_ID	NA	NA	0.154
rider_num_rides	NA	NA	NA
appt_duration	0.154	NA	NA

Table 8: Matrix table of correlation coefficient

Correlation Plot