Exploratory Data Analysis Report

2022-02-17

- Exploratory Data analysis (EDA)
  - 1. Overview of the data
  - 2. Summary of numerical variables
  - 3. Distributions of numerical variables
  - 4. Summary of categorical variables
  - 5. Distributions of categorical variables
    - Bar plots for all categorical variables

Exploratory Data analysis (EDA)

Analyzing the data sets to summarize their main characteristics of variables, often with visual graphs, without using a statistical model.

1. Overview of the data

Understanding the dimensions of the dataset, variable names, overall missing summary and data types of each variables

```
# Overview of the data
ExpData(data=data,type=1)
# Structure of the data
ExpData(data=data,type=2)
```

Overview of the data

Descriptions <chr>	Value <chr>
Sample size (nrow)	12487
No. of variables (ncol)	14
No. of numeric/interger variables	2
No. of factor variables	0
No. of text variables	12
No. of logical variables	0
No. of identifier variables	0
No. of date variables	0
No. of zero variance variables (uniform)	0
%. of variables having complete cases	35.71% (5)

1-10 of 13 rows	Previous **1** 2 Next

Structure of the data

Index <dbl>	Variable_Name <chr>	Variable_Type <chr>	Sample... <int>	Missing_Count <int>	Per_of_Missing <dbl>
1	anon_ID	integer	12487	0	0.000
2	assistance_date	character	12487	0	0.000
3	assistance_category	character	12487	0	0.000
4	amount	numeric	12487	0	0.000
5	unit	character	12487	0	0.000
6	CommType	character	4668	7819	0.626
7	Party	character	4675	7812	0.626
8	InitiatedBy	character	4645	7842	0.628
9	Benefit_1	character	4060	8427	0.675
10	Assistance_1	character	4126	8361	0.670

1-10 of 14 rows | 1-6 of 7 columns   Previous **1** 2 Next

2. Summary of numerical variables

Summary of all numeric variables

```
ExpNumStat(data,by="A",gp=NULL,Qnt=seq(0,1,0.1),MesofShape=2,Outlier=TRUE,round=2)
```

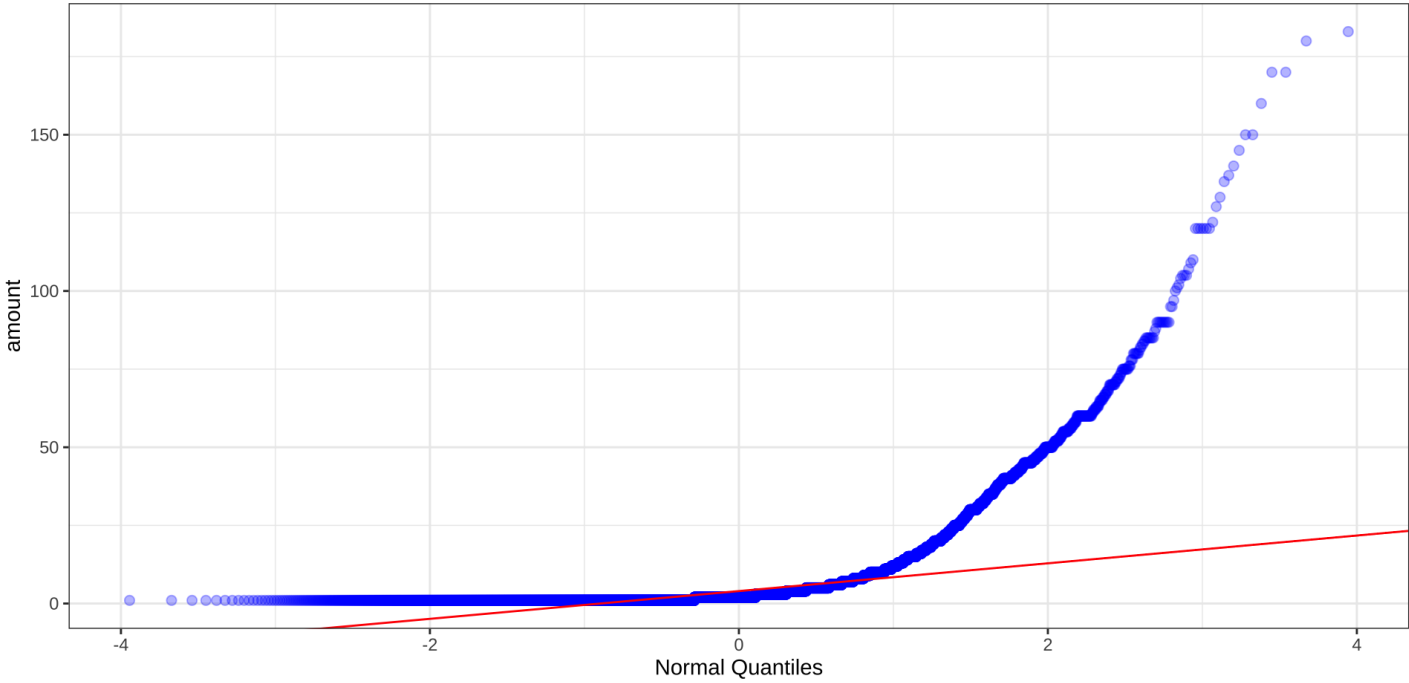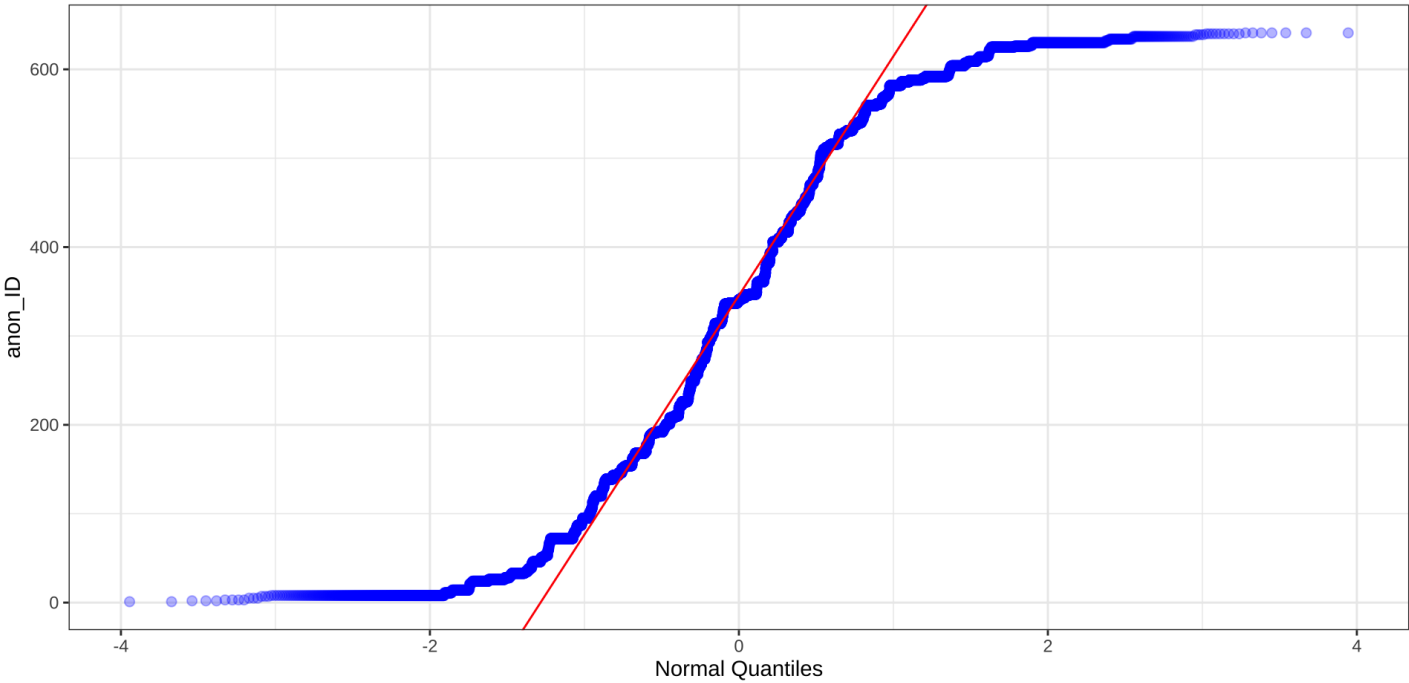| Vname <chr>	Group <chr>	TN <dbl>	n... <dbl>	nZero <dbl>	nPos <dbl>	NegInf <dbl>	PosInf <dbl>	NA_Value <dbl>	Per_of_Missing <dbl>
amount	All	12487	0	0	12487	0	0	0	0
anon_ID	All	12487	0	0	12487	0	0	0	0

2 rows | 1-10 of 34 columns

3. Distributions of numerical variables

Graphical representation of all numeric features

- Quantile-quantile plot(Univariate)

```
ExpOutQQ(data,nlim=4,fname=NULL,Page=c(2,2),sample=sn)
```
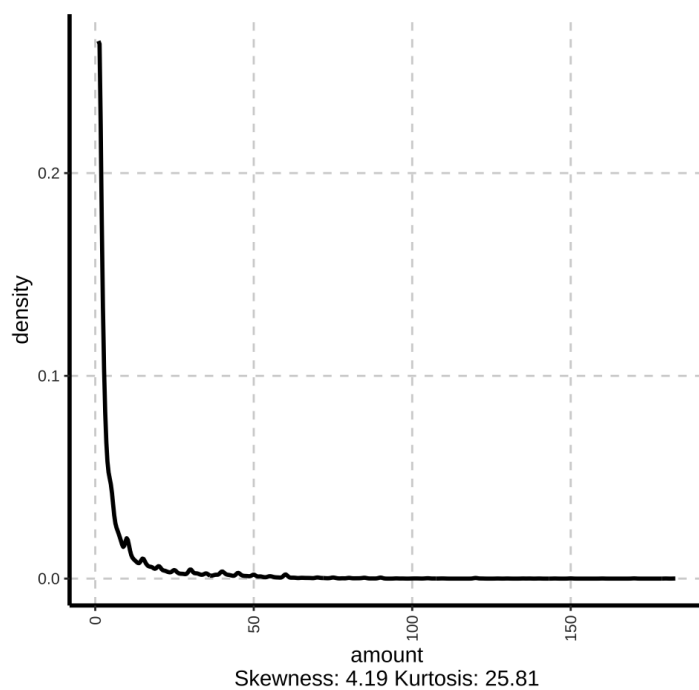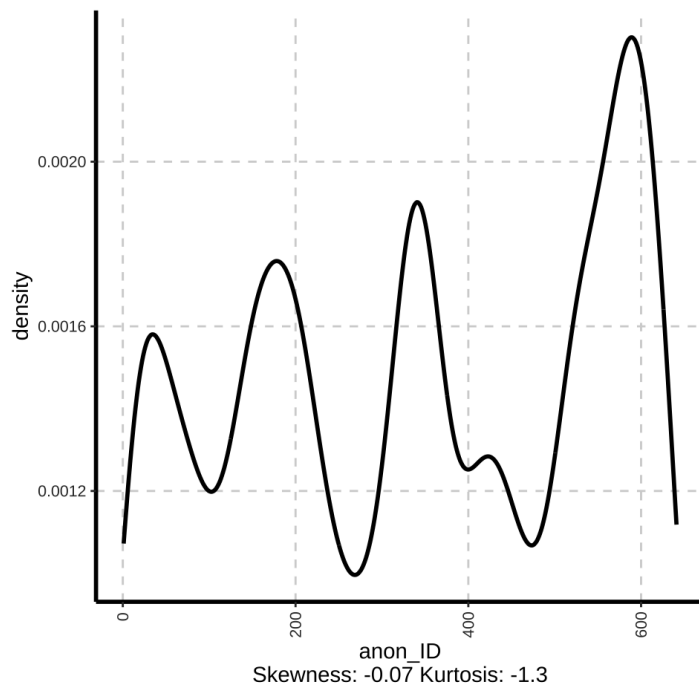
```
## $`0`
```

page 1 of 1

- Density plot (Univariate)

```
ExpNumViz(data,target=NULL,type=1,nlim=10,fname=NULL,col=NULL,Page=c(2,2),theme=theme,sa
mple=sn)
```
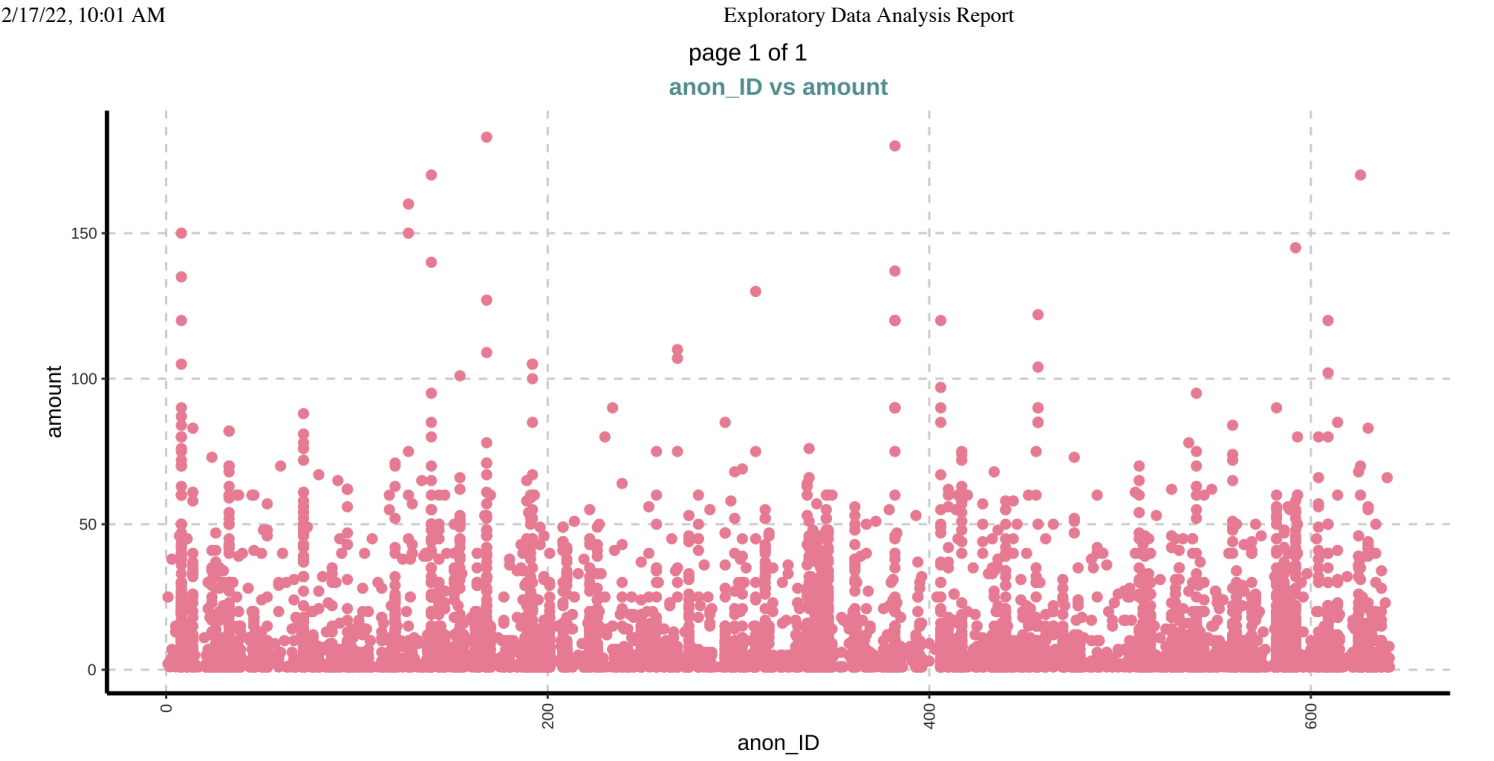
```
## $`0`
```

anon_ID
Skewness: -0.07 Kurtosis: -1.3

amount
Skewness: 4.19 Kurtosis: 25.81

- Scatter plot (Bivariate)

```
ExpNumViz(data,Page=c(2,1),sample=sn,theme=theme,scatter=TRUE)
```

```
## $`0`
```

anon_ID vs amount

4. Summary of categorical variables

Summary of categorical variables

- frequency for all categorical independent variables

```
ExpCTable(data,Target=NULL,margin=1,clim=10,nlim=5,round=2,bin=NULL,per=T)
```

Variable	Valid	Frequency	Percent	C
<chr>	<chr>	<dbl>	<dbl>	
assistance_category	Community	757	6.06	

Variable <chr>	Valid <chr>	Frequency <dbl>	Percent <dbl>	C
assistance_category	ElderNet Office	47	0.38	
assistance_category	Home Visit	766	6.13	
assistance_category	Internal Discussions	233	1.87	
assistance_category	Remote Contact w/o client	1948	15.60	
assistance_category	Remote Contact with client (phone, email)	8736	69.96	
assistance_category	TOTAL	12487	*NA*	
unit	Dollars	10	0.08	
unit	Minutes	12475	99.90	
unit	Pounds	2	0.02	

1-10 of 52 rows   Previous **1** 2 3 4 5 6 Next

NA *is Not Applicable*

5. Distributions of categorical variables

- Bar plots for all categorical variables

Bar plots for all categorical variables

Bar plot with vertical or horizontal bars for all categorical variables

```
ExpCatViz(data,target=NULL,fname=NULL,clim=10,margin=2,theme=theme,Page = c(2,2),sample=
sc)
```

```
## $`0`
```

assistance_category

CommType

unit

InitiatedBy

Assistance_2

Assistance_3