Data Quality report

-UCD Student number 22203536, Summer 2023



**Dataset description**

The dataset being reported is the 2019 Yellow Taxi Trip Dataset. Available at

<https://data.cityofnewyork.us/Transportation/2019-Yellow-Taxi-Trip-Data/2upf-qytp>

This is a large dataset with over 84 million rows that covers information about taxi trips such as pickup taxi zone, dropoff taxi zone, pickup and dropoff times as well as various charges.

A full data dictionary is available at <https://data.cityofnewyork.us/Transportation/2019-Yellow-Taxi-Trip-Data/2upf-qytp> under attachments. Below are the features in the dataset.

VendorID -TPEP provider

tpep\_pickup\_datetime - date and time metre is engaged

tpep\_dropoff\_datetime - date and time metre is disengaged

passenger\_count - number of passengers (driver entered value)

trip\_distance - distance in miles on the taximeter

RatecodeID - rate code in effect at the end of the trip

store\_and\_fwd\_flag - whether data was stored in vehicle before sending to vendor

PULocationID - taxi zone of pickup

DOLocationID - taxi zone of dropoff

payment\_type - how the passenger paid for the trip

fare\_amount - fare in dollars (base)

extra - 50 cent and 1 dollar rush hour and overnight charges

mta\_tax - 50 cent MTA tax

tip\_amount - tip in dollars

tolls\_amount - amount of tolls in dollars

improvement\_surcharge - 30 cent surcharge

total\_amount - total fare including all charges

congestion\_surcharge - charge for high levels of congestion \*(not specified in the data dictionary)

**General data quality**

This analysis is based off of the python notebook titled DataQualityReport.ipynb.

The general data quality of the dataset seems to be fine with some exceptions. In particular the dataset contains a large number of outliers for many of the features. Further analysis is recorded below.

**Missing rows**

VendorID 163435

tpep\_pickup\_datetime 0

tpep\_dropoff\_datetime 0

passenger\_count 163435

trip\_distance 0

RatecodeID 163435

store\_and\_fwd\_flag 163435

PULocationID 0

DOLocationID 0

payment\_type 163435

fare\_amount 0

extra 0

mta\_tax 0

tip\_amount 0

tolls\_amount 0

improvement\_surcharge 0

total\_amount 0

congestion\_surcharge 4585780

Most features have no missing values and the features that do have missing values only have a negligible percentage of rows missing 163435 rows are missing for 5 features. But congestion surcharge has 4585780 rows missing which is substantially higher than other features but still equates to only about 5% of the data.

**Duplicate rows**

There are almost no duplicate rows in this dataset. My analysis only revealed 4 duplicated rows

**Negative values**

Negative values are not coherent for any of the features in this dataset. You cannot have negative distances times or charges. There is no indication from the data dictionary that the negative monetary values could equate to refunds. The number of negative values for each feature are listed below. Most likely the - sign was incorrectly added and the absolute values of these numbers represent the true values.

payment\_type: 0

VendorID: 0

passenger\_count: 0

trip\_distance: 6379

RatecodeID: 0

PULocationID: 0

DOLocationID: 0

fare\_amount: 138930

extra: 69385

mta\_tax: 136433

tip\_amount: 763

tolls\_amount: 3101

improvement\_surcharge: 138742

total\_amount: 138908

congestion\_surcharge: 118992

**Variance level**

Variance level refers to how dissimilar the rows of a particular feature are. If the variance is low it means that there is very high similarity among the rows of a feature. Such features are not informative for machine learning models.

feature: payment\_type

variance level: 0.22211193917582667

feature: VendorID

variance level: 0.24842182853239686

feature: passenger\_count

variance level: 1.46383851372644

feature: trip\_distance

variance level: 21.77741744592158

feature: RatecodeID

variance level: 0.36680173677158295

feature: PULocationID

variance level: 4338.166601561239

feature: DOLocationID

variance level: 4796.240597213574

feature: fare\_amount

variance level: 21694.971572884442

feature: extra

variance level: 1.5924042900291648

feature: mta\_tax

variance level: 0.003055942247459988

feature: tip\_amount

variance level: 256.8065570835164

feature: tolls\_amount

variance level: 2.950818985336107

feature: improvement\_surcharge

variance level: 0.0006530394824207707

feature: total\_amount

variance level: 25353.634913588594

feature: congestion\_surcharge

variance level: 0.48471581789780427

The variance levels for mta\_tax and improvement surcharge are most notable very low and so the rows have high similarity level. As such we may choose to drop these features later.

**Collinearity**

A feature is said to be collinear with another feature if there is a strong positive or negative relationship with the other feature. This means that the data for one feature might be captured mostly in the other feature and as such having both features to train the model may not be a good idea. A correlation matrix was made in the jupyter notebook referenced above and it was found that fare\_amount and total\_amount were highly collinear with a score of 0.997387. This makes sense considering that the amount of base fare plus a number of standard charges and tips equates to the total fare. As such we may choose later to drop one of these features.

**Incoherent values**

Many of the values in the dataset are not coherent. For example there are taxi fares below the minimum charge for taxi journeys which was 2.5 dollars in 2019[1]. The data dictionary does allow for some journeys to have no charge under payment\_type but payments under 2.50 and more than 0 should not be there. There are journeys with passenger counts of 0 and journeys of passenger counts greater than the number of car seats even if you allow for the NYC law that allows for lapsitting[2]. Further details will be provided below in the feature analysis.

**Outliers**

Outliers were calculated using the interquartile range and are listed below

the number of outliers in passenger\_count: 6822193

the number of outliers in trip\_distance: 8616584

the number of outliers in fare\_amount: 7245537

the number of outliers in extra: 115547

the number of outliers in mta\_tax: 526965

the number of outliers in tip\_amount: 3377959

the number of outliers in tolls\_amount: 4674154

the number of outliers in improvement\_surcharge: 166209

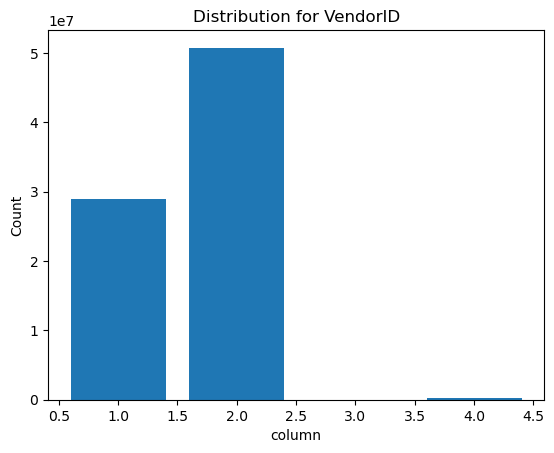
the number of outliers in total\_amount: 7254872

the number of outliers in congestion\_surcharge: 599687

This represents the single biggest data quality issue since there are millions of outliers for many of the features and so removing them would lead to a major loss in data volume. Outliers may represent reasonable values for some features. For example if there are outliers in tip\_amount this might be reasonable since some passengers might tip more generously than others. As such I explored the minimum and maximum values for many of these features like trip\_distance had a max value of 19130.18 miles. Considering the size of New York this is not a reasonable value so some of these outliers may represent poor data quality and not reasonable values.

**Features**

**VendorID**: A code indicating the TPEP provider that provided the record. 1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.

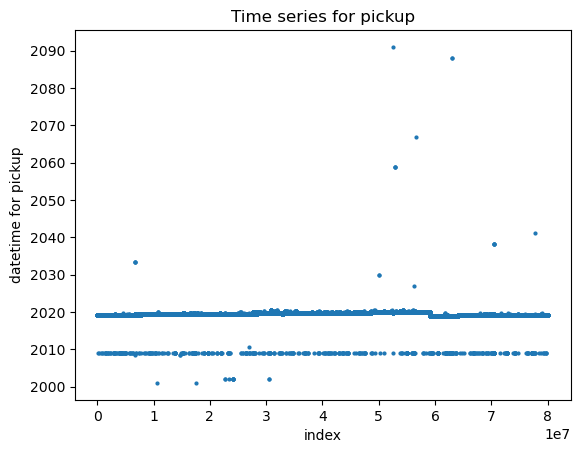


The different values for VendorID are 2. 1. 4. Nan according to our analysis in the notebook. The modal group is 2 as per the histogram. But the data dictionary only described 2 possible vendors. Thus the entries as 4 may be a mistake or the data dictionary may be incomplete (since the number of features in the provided data dictionary is one less than the number of features in the dataset). The vast majority of the rows seem to be either one or 2 with a tiny minority being 4 as per the diagram. 163435 values are missing but this is only about 0.2% of the dataset. There are no negative values for this feature.

Variance level for this feature is 0.24842182853239686. Noteworthy though is that this feature has a strong negative correlation with extra which might lead us to consider removing one of them as they are collinear.

**Tpep\_pickup\_datetime**

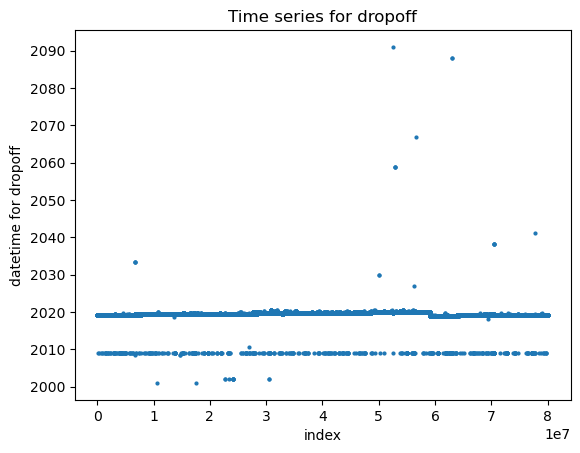
The date and time when the metre was engaged.



Majority of the data seems to be for 2019 as expected since it's a 2019 dataset. But some data points for other years are also available. In particular, 2009 seems to be a year with a lot of data points with 355 data points according to the analysis . These are most likely mistaken entries that were meant for 2019. We will later consider removing or imputing the mistaken entries. There are no missing values for this feature. There are 987 values for which the pickup time is after the dropoff time. These are most likely entries where the entries for pickup and dropoff are switched since logically the dropoff time must be later.

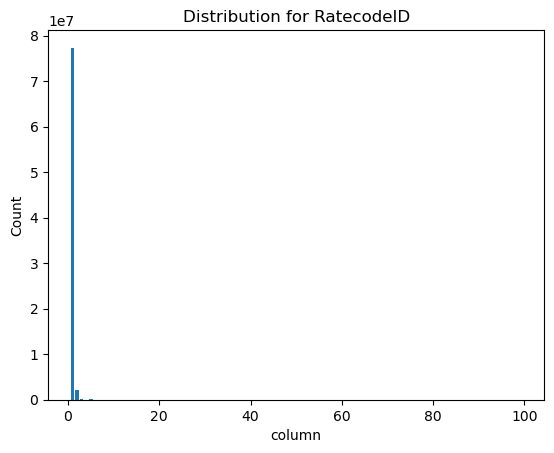
**Tpep\_dropoff\_datetime**

Defined as the date and time that the metre was disengaged.



Again we see the same issue of values for 2009 being entered in addition to 2019 with 420 values for 2009 as well as values for other years. More than likely mistakes where 2009 was meant to refer to 2019. Again there are no missing values for this feature and we as mentioned before have an issue of a small number of dropoff times being before the pickup times. There are 987 values for which the pickup time is after the dropoff time. Again this is most likely due to the times being switched accidentally and we will later consider switching the times back. There are no negative values for this feature. The variance level is variance level: 0.36680173677158295.

**RatecodeID**

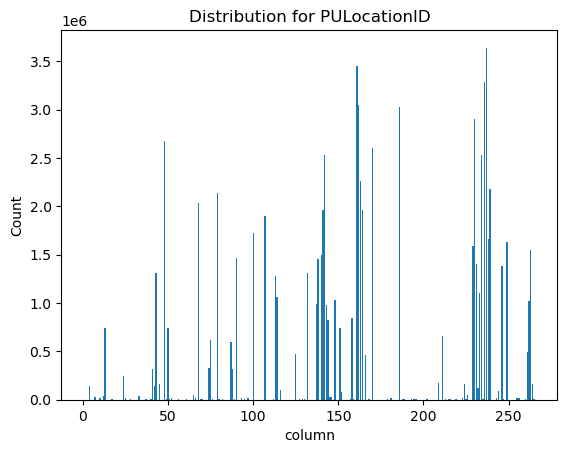


Defined as the final rate code in effect at the end of the trip. 1= Standard rate 2=JFK 3=Newark 4=Nassau or Westchester 5=Negotiated fare 6=Group ride

The modal value is 1. The different values according to the analysis are [ 1. 4. 2. 5. 3. 99. 6. Nan]

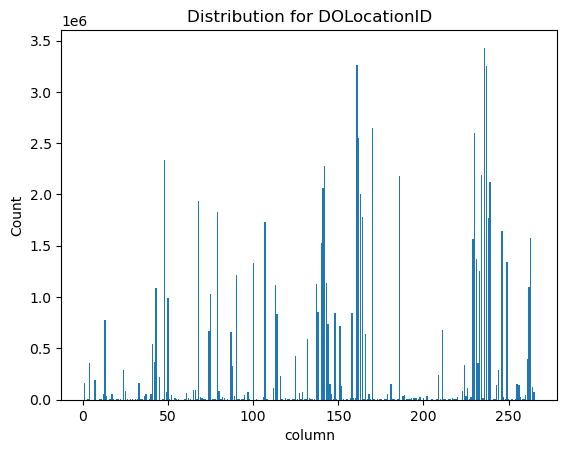
The values that are 99 are most likely a mistake and will consider removing or imputing these values. 163435 missing values. Since it has the same number of missing values as many other features i.e VendorID also has exactly 163435 rows missing, it's quite likely that these represent rows with many missing features. We will consider removing these rows.

**PULocationID**

****

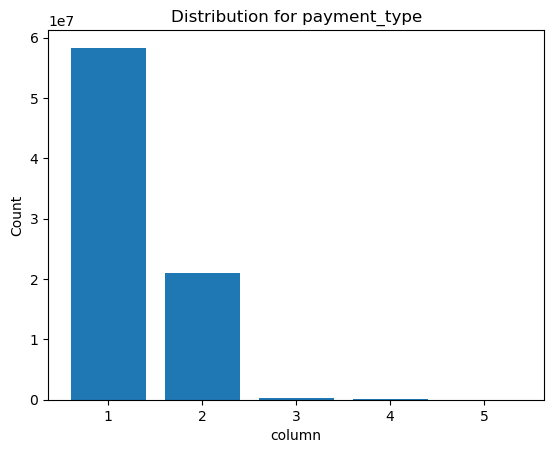
Defined as the TLC Taxi Zone in which the taximeter was engaged The mode is zone 237. Looking at the diagram there seem to be some zones that are far busier than others. This is reasonable as the taxi zones will vary by many factors such as size. population etc that will affect busyness. There are no missing or negative values for this feature which is a good sign for data quality. Variance level is extremely high 4338.166601561239 because there are lots of different values for this feature.

**DOLocationID**

****

Defined as the TLC Taxi Zone in which the taximeter was disengaged The mode is zone 236. Looking at the diagram there seem to be some zones that are far busier than others. This is reasonable as the taxi zones will vary by many factors such as size. population etc that will affect busyness. There are no missing or negative values for this feature which is a good sign for data quality. Variance level is extremely high 44796.240597213574 because there are lots of different values for this feature.

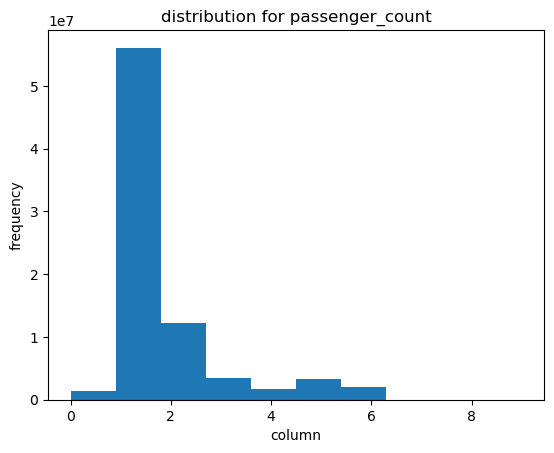
**Payment\_type**



Defined in the data dictionary as a numeric code signifying how the passenger paid for the trip. 1= Credit card 2= Cash 3= No charge 4= Dispute 5= Unknown 6= Voided trip. The most popular payment types are credit cards followed by cash and other values are predictably less common. There are no negative values for this feature. Because most of the values are 1 or 2 the variance level is very low at 0.22211193917582667. The different values in the dataset are

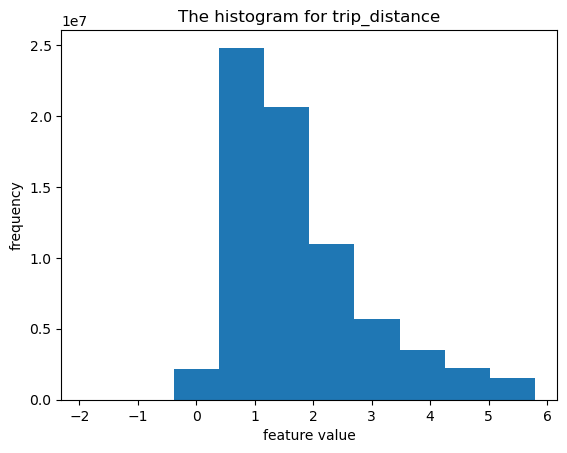
[ 1. 2. 3. 4. 5. Nan] which are coherent with the data dictionary. 163435 missing values. Since its the same number of missing values as many other features i.e VendorID also has exactly 163435 rows missing, it's quite likely that these represent rows with many missing features. We will consider removing these rows.

**Passenger\_count**



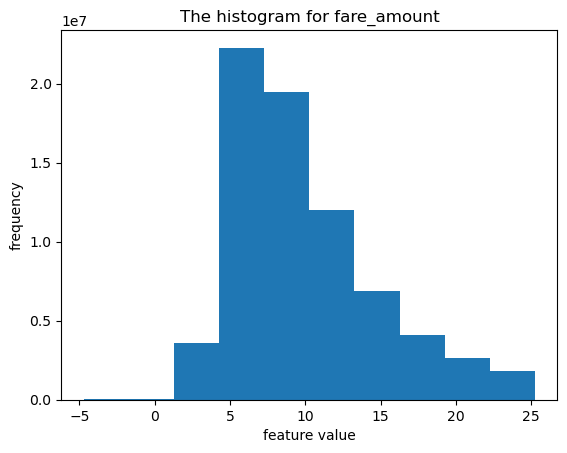
This feature represents the number of passengers and is a driver entered value. It being a driver entered value might explain why passenger count is 0 for 1.4 million rows. The driver may simply have forgotten to take note of the number of passengers. There are values higher than 6 that are most likely mistakes since the larger NYC yellow taxis only seat 5 and seat 6 if a child sits on the lap of an adult which is permitted. The mean passenger count is approximately 1. The median is 1 and the mode is 1. This suggests that most taxi journeys in Manhattan are single journeys. There are no negative values for this feature but there are 163435 missing values which are likely the same missing values across many features since other features also have exactly 163435 missing values. The number of outliers is large at 6822193 which represents about 8.5% of the data. The variance level is 1.46 suggesting moderate levels of variation in this feature. The data quality is fair.

**Trip Distance**

****

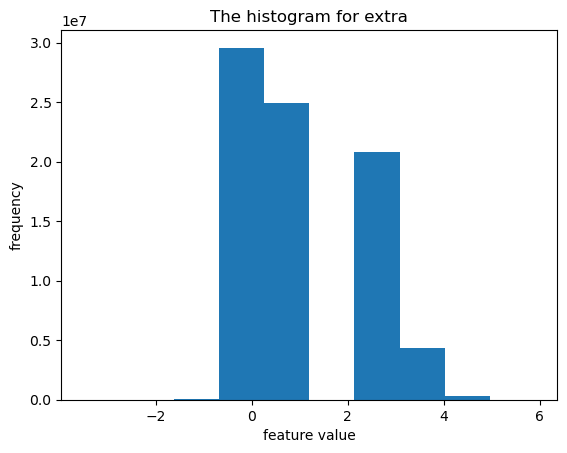
This feature represents the elapsed distance as reported by the taximeter in miles . As per the histogram the vast majority of journeys (when outliers are excluded) are below 6 miles. The mean is 2.8 miles and the median is 1.6 miles. The modal trip is 0.9 miles. There are 6300 negative values for this feature, most likely incorrect entries since distances cannot be negative. There are no missing values for this feature but there are a high number of outliers at over 8.6 million which is more than 10% of the dataset. The variance level is 21.78 which is high suggesting a large variation in trip distances. The data quality is fair overall.

**Fare\_amount**

****

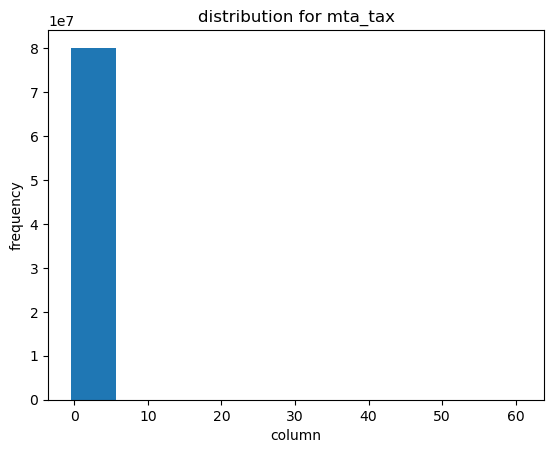
Defined as the time and distance fare calculated by the meter. The majority of the fares as per the diagram are below 25$. There is a sharp cutoff between 0 and 5 most likely caused by the 2.50$ base charge[1]. In2019 there was a base charge of 2.50$ so fares below this amount are not valid.There are 22057 fares between 0 and 2.50 so these are most likely mistakes. The mean fare is 12.65$ and the median is 9.50$. The modal fare is 6$. There are also 138930 negative fares which are not valid as the data dictionary makes no mention of refunds. There's a large number of outliers at over 7 million and a large variance of 21694 which indicates a wide range in the fare. The data quality is fair overall.

**Extra:**

****

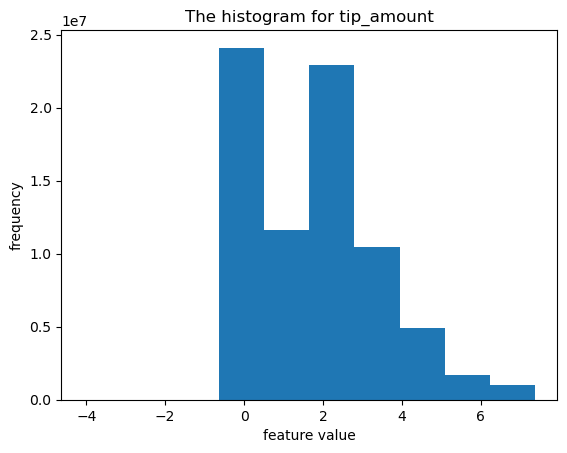
This feature represents extra charges and surcharges. The mode is 0$ suggesting many passengers do not pay this charge. The mean and median are 50 cent and 1$ respectively. There are 69385 negative values which are likely mistakes and should be their positive counterparts. The variance level is 1.59 which is a moderate level of variation however there is a strong negative correlation of -0.8 between this feature and VendorID so we may consider removing it.

**MTA\_tax**

****

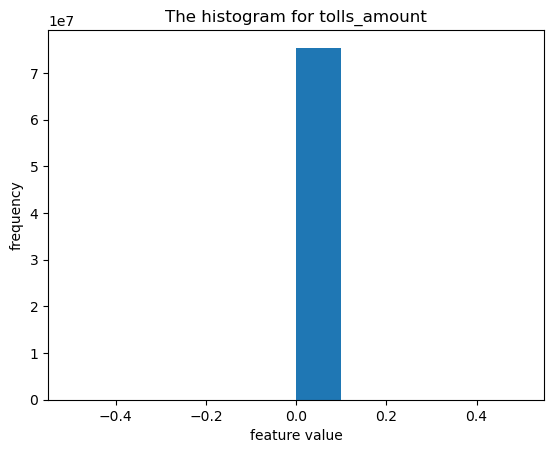
This is a standard 50 cent charge based on metered rate in use. The variance of this feature is very low at 0.003 so we will consider dropping this feature.

**Tip\_amount**

****

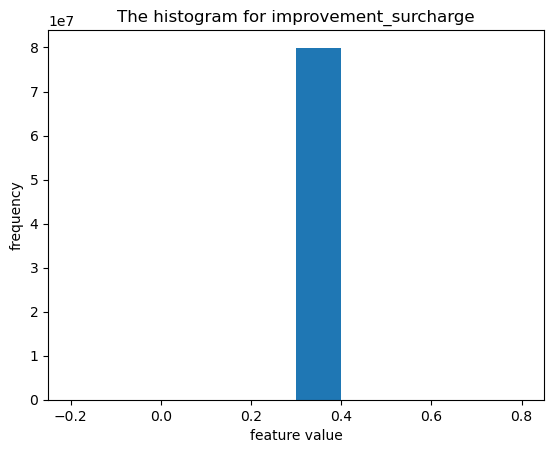
This field is tip amount in dollars not including cash tips. The range of most tips seems to be fom 0-6$ according to the histogram. The mean is 2.17$ and the median is 1.86$. The mode is 0$ suggesting many passengers may not tip. There are no missing values and only 763 negative values which are invalid as you cannot have negative tips. The feature has a reasonably high number of outliers of 3.3 million representing over 4% of the dataset. The variance level is 256.81 which indicates that tips are highly varied. The data quality is fair overall.

**Tolls\_amount**

****

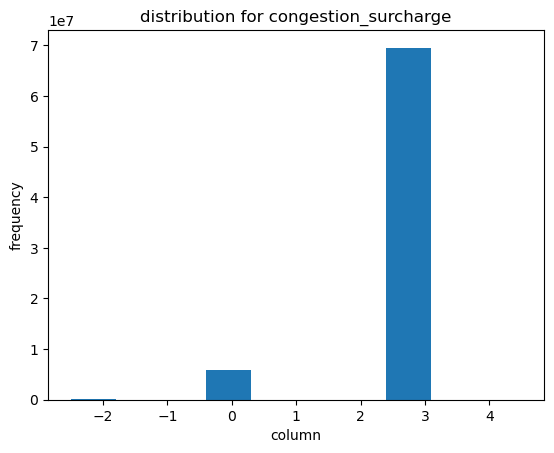
This feature represents the total amount of tolls paid on the trip. The mean is 0.37$ and the median is 0$ suggesting most passengers may not pay tolls. There are no missing values but there are quite a number of outliers at 4.6 million representing over almost 6% of the data. The variance level is 2.95 indicating only moderate variation in tolls so the data quality overall is fair.

**Improvement\_surcharge**

****

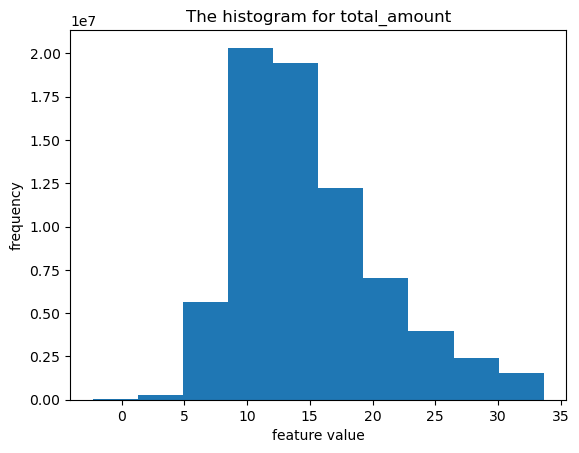
This feature refers to the 30 cent improvement surcharge at the flag drop. Since its a standard charge of 30 cent all the values fall around 0.3 on the histogram. The mean median and mode are all 30 cent and the variation is extremely low at 0.00065 which means there is little variation in the data at all. As such we will consider dropping this feature.

**Congestion\_surcharge**

****

Not defined in the data dictionary but given the unique values we can presume that its a charge based on the different levels of congestion [ 2.5 0. -2.5 0.75 2.75 2.25 -0.75 0.8 1. 0.5 nan -1.5 1.5 3. 4.5 ]. Variance level is 0.4847 which is a moderate level of variation. The mean is 2.29$ and the median is 2.50$. There are just under 6 million outliers for this feature which represents about 7.5% of the dataset. There are quite a lot of missing rows too at 4.6 million rows missing the highest of any feature. There are just under 119000$ negative values which are likely errors as negative values for congestion would be highly unusual.

**Total\_amount**

****

After removing outliers to better visualise this feature most of the values fall between 5 and 35$. These are reasonable figures. The mean is 18.50$ and the median is 14.30$ which look appropriate based on the histogram. However this feature is extremely highly correlated with fare\_amount at 0.997387

Correlation value. As such we will consider removing it. The high positive correlation makes sense since this feature is just the sum of fare\_amount and some other charges to the customer.

**References**

1. Bloomberg (2022) 'NYC Taxi Cab Fares to Rise 23% in First Increase Since 2012', Bloomberg. Available at:<https://www.bloomberg.com/news/articles/2022-11-15/nyc-taxi-cab-fares-to-rise-23-in-first-increase-since-2012> (Accessed: 7 July 2023).
2. NYC.gov (n.d.) 'TLC Rules', NYC.gov. Available at:<https://www.nyc.gov/site/tlc/about/tlc-rules.page> (Accessed: 7 July 2023).