1 Numbers of errors

1.1 VendorID

By the summary the minimum value is 1 and the maximum value is 2. There are 1070 rows with value 1, and 3930 rows with value 1. Therefore nothing to be dropped here and no errors.

1	2
1070	3930

1.2 lpep_pickup_datetime Lpep_dropoff_datetime

Needs further investigation as datetime.

1.3 Store_and_fwd_flag

By print the table I can observe that there are 4987 rows with value No, and 13 rows with value Yes. This leaves no errors.

Y	N
13	4987

1.4 RateCodeID

By printing the table I can observe that there are 5000 valid rows.

1	2	3	4	5
4866	23	3	1	107

1.5 Pickup_longitude

This variable has a minimum value of -75.12, and a maximum value of 0.00. This is a huge, and not limited to Manhattan value. There are 16 rows with 0.00 incorrectly set as a pickup longitude.

1.6 Pickup_latitude

This variable has a minimum value of 0.00, and a maximum value of 40.90. This is a huge, and not limited to Manhattan value. There are 16 rows with 0.00 incorrectly set as a pickup longitude. This leaves a question, are these the same cases as the pickup longitude.

1.7 Dropoff_longitude

This variable has a minimum value of -75.12, and a maximum value of 0.00. This is a huge, and not limited to Manhattan value. There are 16 rows with 0.00 incorrectly set as a pickup longitude. This leaves a question, are these the same cases as the pickup longitude.

1.8 Dropoff_latitude

This variable has a minimum value of 0.00, and a maximum value of 40.93. This is a huge, and not limited to Manhattan value. There are 16 rows with 0.00 incorrectly set as a pickup longitude. This leaves a question, are these the same cases as the pickup longitude.

1.9 Passenger_count

This variable has a minimum number of 0, with a maximum of 6. A journey shouldn't have zero passengers, and this is clearly a mistake.

1.10 Trip_distance

There are 83 0 mile journeys which are errors.

zero values	non-zero values
83	4917

1.11 Fare_amount

The minimum value is -10, the maximum value is 124.50. There are 19 cases where the fare amount is less than or equal to 0. These are errors.

1.12 Extra

The minimum value is -1, the maximum value is 1. There are 5 cases lower than 0.

1.13 MTA_tax

The minimum value is -0.5, the maximum value is 0.5. There are 3 cases lower than 0.

1.14 Tip_amount

The minimum value is 0, the maximum value is 51.00. There is nothing missing here although there are suspiciously high tip amounts.

1.15 Tolls_amount

The minimum value is 0, the maximum value is 20.00. There is nothing suspicious here.

1.16 Ehail_fee

The whole column is NA, so we should simply drop this column

1.17 improvement_surcharge

The improvement surcharge ranges from -0.3 to 0.3. Anything below 0 is an error.

1.18 Total_amount

The minimum value for total amount is -10, the maximum is 125.30. Anything below 0 is an error, and there are 5 such values.

1.19 Payment_type

There is nothing out of the ordinary here.

1	2	3	4
2528	2432	24	16

1.20 Trip_type

There is nothing out of the ordinary here.

1	2
4895	105

1.21 total_time

The minimum value is 0, the maximum value is 86141. These are erroneously high and low values. The zeros are errors, the high rows are outliers and should be treated as such. 86141 is just under a full day, there also shouldn't be 0 second journeys. drop 0 or below, there are also high outliers

1.22 Missing Data

The only missing data is in the ehail fees column, the entire column is NA, and so we cannot really use this.

1.23 Errors

Pickup/Dropoff Longitude/Latitude contribute 16 errors with 0 values. Passenger Count has 0 values, this is an error. Trip distance has 83 0 mile trips. Fare amount has 19 less than zero values. Extra has 5 cases less than 0. MTA tax has 3 cases less than 0. Improvement surcharge has negative values, these are errors. Total Amount, 5 values less than 0. Total time, 0 and 86141 is clearly an error.

1.24 Number of Errors

0	1	2	3	4	5	6
4897	72	12	7	7	4	1

1.25 Errors per variable

VendorID	0
lpep_pickup_datetime	?
Lpep_dropoff_datetime	?
Store_and_fwd_flag	0
RateCodeID	0
Pickup_longitude	16
Pickup_latitude	16
Dropoff_longitude	16
Dropoff_latitude	16
Passenger_count	1
Trip_distance	83
Fare_amount	5
Extra	2
MTA_tax	3
Tip_amount	0
Tolls_amount	0
Ehail_fee	0
improvement_surcharge	3
Total_amount	5
Payment_type	0
Trip_type	0
total_time	0

Or in order of least trustworthy variables.

Trip_distance	83
Pickup_longitude	16
Pickup_latitude	16
Dropoff_longitude	16
Dropoff_latitude	16
Fare_amount	5
Total_amount	5
improvement_surcharge	3
MTA_tax	3
Extra	2
Passenger_count	1
VendorID	0
Store_and_fwd_flag	0
RateCodeID	0
Tip_amount	0
Tolls_amount	0
Ehail_fee	0
Payment_type	0
Trip_type	0
total_time	0
lpep_pickup_datetime	?
Lpep_dropoff_datetime	?

1.26 Outliers

Tip Amount has outliers, though I don't know that these should be ignored. The 95th Quartile is 5. Total Time, There are extreme outliers which need to be taken into account. the 95th Quartile is 1989. The 2nd Quartile is 60. I would class anything satisfying these as outliers.

1.27 Number of Outliers

0	1	2
4592	323	85

1.28 Number of Missing

The only missing values are for Ehail fees. I haven't added them since they are simply all missing.

2 Conclusion

I would suggest that any row that has an outlier or error isn't ready to be used in this model. Additionally i have not investigated the pickup and drop off times which may be invalid as well and require further investigation.