Total data

```
Total rows: 527311461
```

Schema on whole data

```
root
 |-- store_and_fwd_flag: string (nullable = true)
 |-- passenger_count: double (nullable = true)
 |-- extra: double (nullable = true)
 |-- tip_amount: double (nullable = true)
 |-- fare_amount: double (nullable = true)
 |-- airport_fee: double (nullable = true)
 |-- total_amount: double (nullable = true)
 |-- DOLocationID: long (nullable = true)
 |-- mta_tax: double (nullable = true)
 |-- trip_distance: double (nullable = true)
 |-- PULocationID: long (nullable = true)
 |-- VendorID: long (nullable = true)
 |-- tpep_pickup_datetime: timestamp (nullable = true)
 |-- tpep_dropoff_datetime: timestamp (nullable = true)
 |-- improvement_surcharge: double (nullable = true)
 |-- payment_type: long (nullable = true)
 |-- congestion_surcharge: double (nullable = true)
 |-- RatecodeID: double (nullable = true)
 |-- tolls_amount: double (nullable = true)
 |-- year: integer (nullable = true)
```

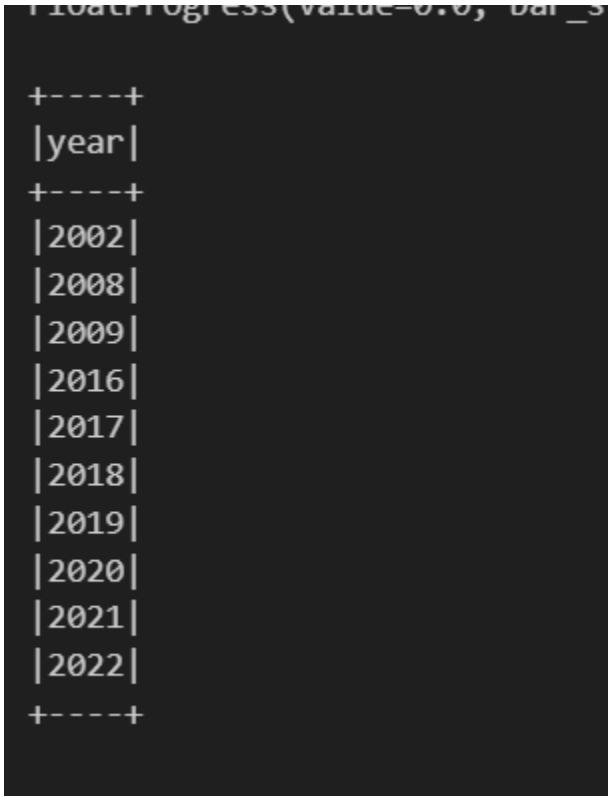
Duplicate rows on whole data

```
Duplicate Rows: 13219
```

Null value on whole data

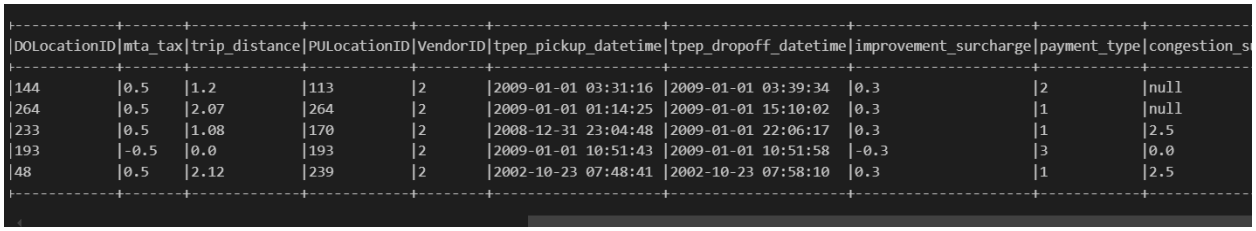
store_and_fwd_flag → 4,159,792
passenger_count → 4,159,792
extra → 0
tip_amount → 0
fare_amount → 0
airport_fee → 463,760,743
total_amount → 0
DOLocationID → 0
mta_tax → 0
trip_distance → 0
PULocationID → 0
VendorID → 0
tpep_pickup_datetime → 0
tpep_dropoff_datetime → 0
improvement_surcharge → 0
payment_type → 0
congestion_surcharge → 356,461,045
RatecodeID → 4,159,792
tolls_amount → 0
year → 0

Sample data (Random)



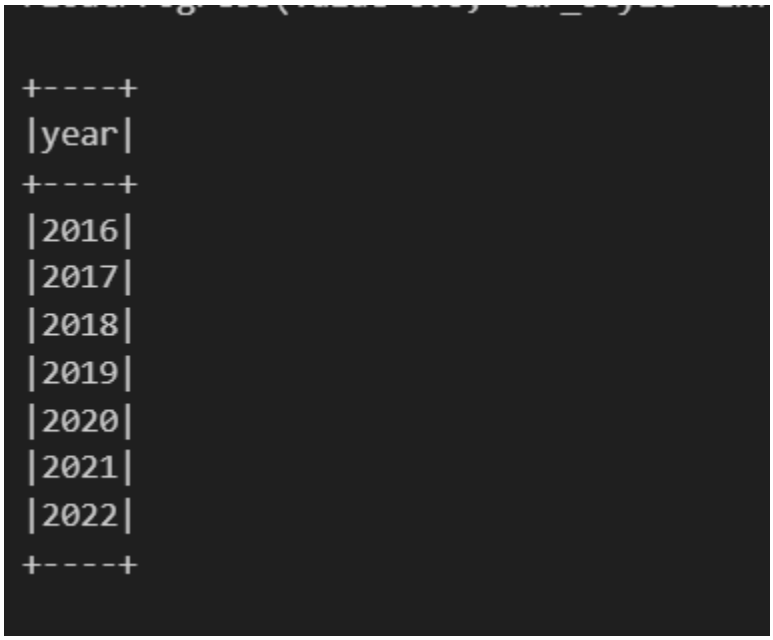
```
+----+
|year|
+----+
|2002|
|2008|
|2009|
|2016|
|2017|
|2018|
|2019|
|2020|
|2021|
|2022|
+----+
```

Dropped Wrong values



DOLocationID	mta_tax	trip_distance	PULocationID	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	improvement_surcharge	payment_type	congestion_s
144	0.5	1.2	113	2	2009-01-01 03:31:16	2009-01-01 03:39:34	0.3	2	null
264	0.5	2.07	264	2	2009-01-01 01:14:25	2009-01-01 15:10:02	0.3	1	null
233	0.5	1.08	170	2	2008-12-31 23:04:48	2009-01-01 22:06:17	0.3	1	2.5
193	-0.5	0.0	193	2	2009-01-01 10:51:43	2009-01-01 10:51:58	-0.3	3	0.0
48	0.5	2.12	239	2	2002-10-23 07:48:41	2002-10-23 07:58:10	0.3	1	2.5

Data from 2016 to 2022



```
+----+
|year|
+----+
|2016|
|2017|
|2018|
|2019|
|2020|
|2021|
|2022|
+----+
```

Finding null values

store_and_fwd_flag → 7859
passenger_count → 7859
extra → 0
tip_amount → 0
fare_amount → 0
airport_fee → 879606
total_amount → 0
DOLocationID → 0
mta_tax → 0
trip_distance → 0
PULocationID → 0
VendorID → 0
tpep_pickup_datetime → 0
tpep_dropoff_datetime → 0
improvement_surcharge → 0
payment_type → 0
congestion_surcharge → 676058
RatecodeID → 7859
tolls_amount → 0
year → 0

Dropped airport_fee and congestion_surcharge ⇒ columns
And ratecodeID , store_and_fwd_flag, passenger_count ⇒ rows

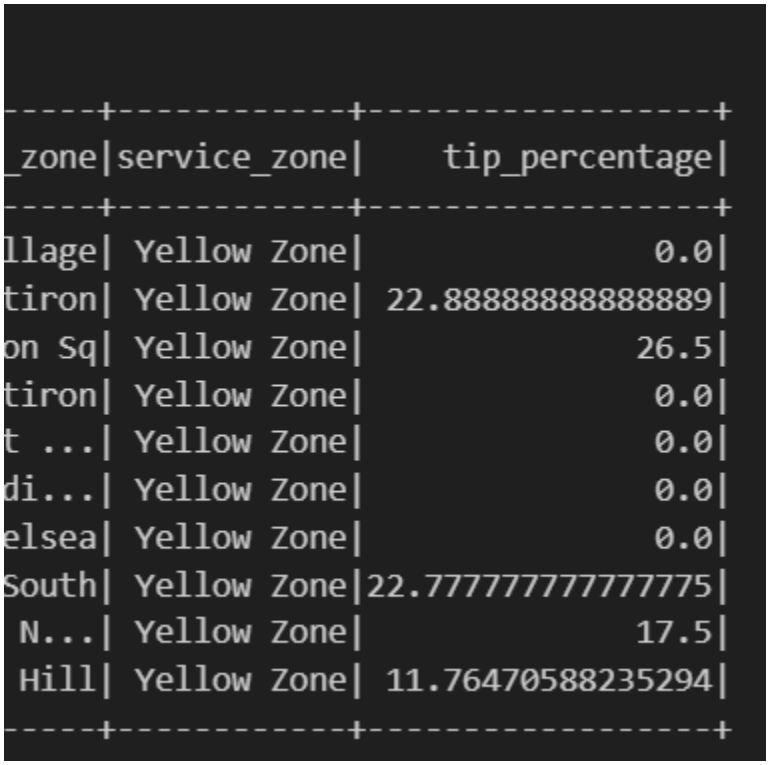
Mapped zone code

PULocationID	Borough	pickup_zone	service_zone
1	EWR	Newark Airport	EWR
2	Queens	Jamaica Bay	Boro Zone
3	Bronx	Allerton/Pelham Gardens	Boro Zone
4	Manhattan	Alphabet City	Yellow Zone
5	Staten Island	Arden Heights	Boro Zone
6	Staten Island	Arrochar/Fort Wadsworth	Boro Zone
7	Queens	Astoria	Boro Zone
8	Queens	Astoria Park	Boro Zone
9	Queens	Auburndale	Boro Zone
10	Queens	Baisley Park	Boro Zone

only showing top 10 rows

s_amount	year	Borough	pickup_zone	service_zone	Borough	dropoff_zone	service_zone
0.0	2016	Manhattan	East Chelsea	Yellow Zone	Manhattan	East Village	Yellow Zone
0.0	2016	Manhattan	SoHo	Yellow Zone	Manhattan	Flatiron	Yellow Zone
0.0	2016	Manhattan	Union Sq	Yellow Zone	Manhattan	Union Sq	Yellow Zone
0.0	2016	Manhattan	Flatiron	Yellow Zone	Manhattan	Flatiron	Yellow Zone
0.0	2016	Manhattan	Hudson Sq	Yellow Zone	Manhattan	Meatpacking/West ...	Yellow Zone

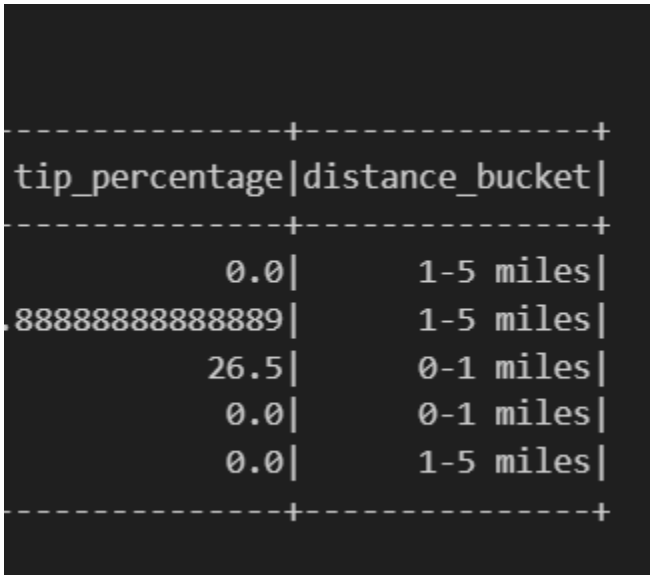
Add new columns tip percentage



```
-----+------------+------------------+
_zone|service_zone|    tip_percentage|
-----+------------+------------------+
llage| Yellow Zone|               0.0|
tiron| Yellow Zone| 22.88888888888889|
on Sq| Yellow Zone|              26.5|
tiron| Yellow Zone|               0.0|
t ...| Yellow Zone|               0.0|
di...| Yellow Zone|               0.0|
elsea| Yellow Zone|               0.0|
South| Yellow Zone|22.777777777777775|
 N...| Yellow Zone|              17.5|
 Hill| Yellow Zone| 11.76470588235294|
-----+------------+------------------+
```

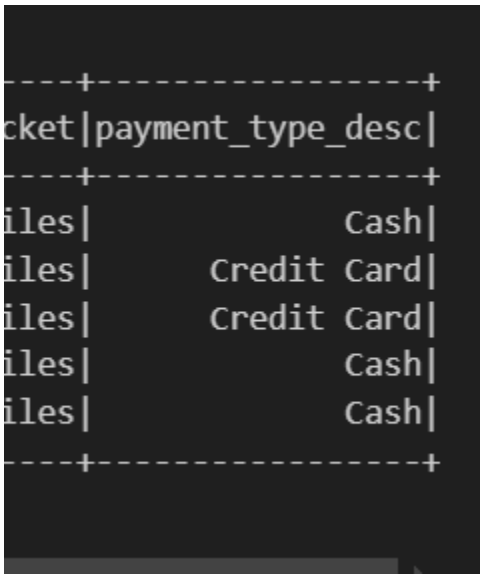
Removed outliers by taking values only of fare amount less then 100 and distance less then 500 miles

Added new columns mile range



```
---------------+---------------+
 tip_percentage|distance_bucket|
---------------+---------------+
            0.0|      1-5 miles|
.88888888888889|      1-5 miles|
           26.5|      0-1 miles|
            0.0|      0-1 miles|
            0.0|      1-5 miles|
---------------+---------------+
```

Added new columns of payment type and mapped with code



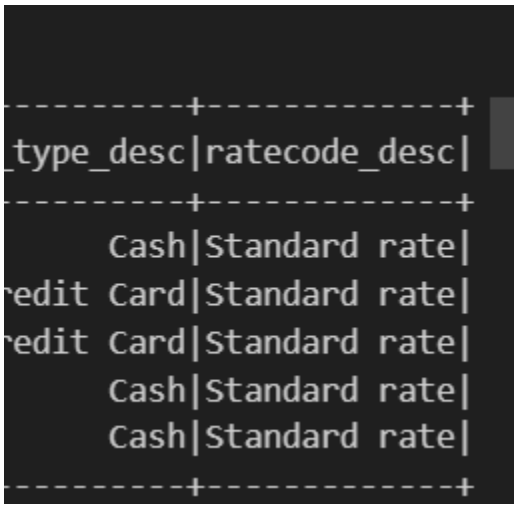
```
----+------------------+
cket|payment_type_desc|
----+------------------+
iles|              Cash|
iles|       Credit Card|
iles|       Credit Card|
iles|              Cash|
iles|              Cash|
----+------------------+
```

Mapped ratecode id

```
+----------+
|RatecodeID|
+----------+
|         1|
|         2|
|         3|
|         4|
|         5|
|         6|
|        99|
+----------+
```

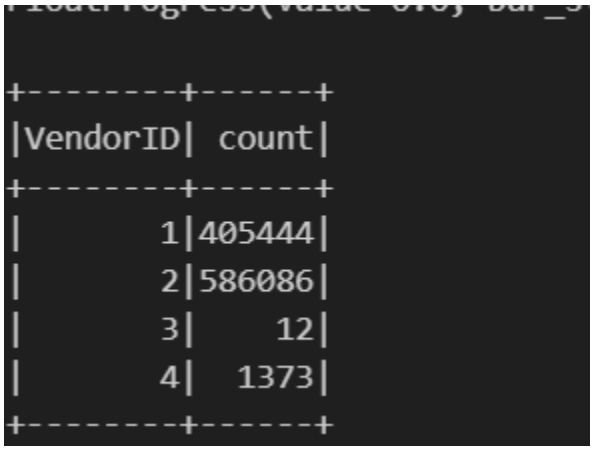
```
from pyspark.sql.functions import when

clean_sample_df = clean_sample_df.withColumn(
    "ratecode_desc",
    when(col("RatecodeID") == 1, "Standard rate")
    .when(col("RatecodeID") == 2, "JFK")
    .when(col("RatecodeID") == 3, "Newark")
    .when(col("RatecodeID") == 4, "Nassau or Westchester")
    .when(col("RatecodeID") == 5, "Negotiated fare")
    .when(col("RatecodeID") == 6, "Group ride")
    .when(col("RatecodeID") == 99, "Unknown")
    .otherwise("Other")
)
```

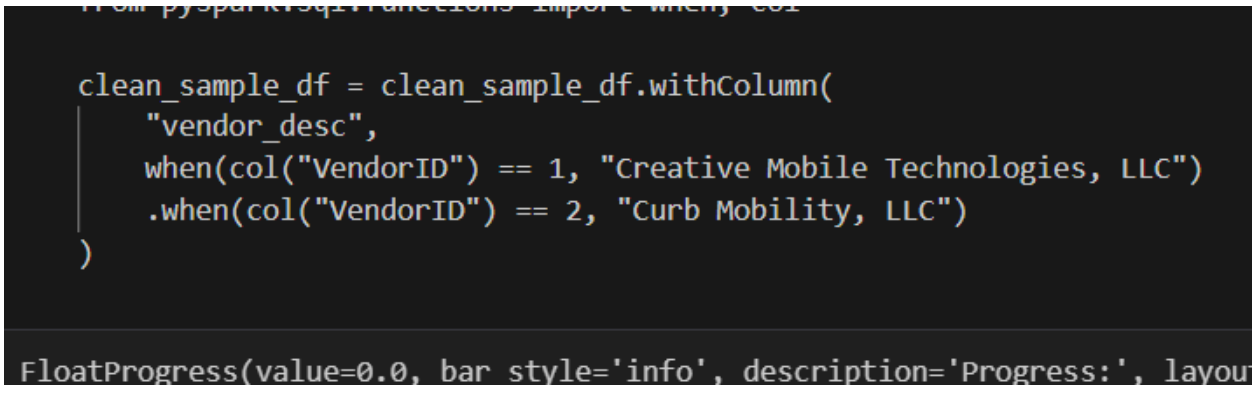


```
---------+-------------+
_type_desc|ratecode_desc|
---------+-------------+
     Cash|Standard rate|
redit Card|Standard rate|
redit Card|Standard rate|
     Cash|Standard rate|
     Cash|Standard rate|
---------+-------------+
```

Mapped vendor id and dropped vendor 3 and 4 because the count was very low



```
+--------+------+
|VendorID| count|
+--------+------+
|       1|405444|
|       2|586086|
|       3|    12|
|       4|  1373|
+--------+------+
```



```
clean_sample_df = clean_sample_df.withColumn(
    "vendor_desc",
    when(col("VendorID") == 1, "Creative Mobile Technologies, LLC")
    .when(col("VendorID") == 2, "Curb Mobility, LLC")
)
```

```
FloatProgress(value=0.0, bar_style='info', description='Progress:', layou
```

```
pyspark

-----+--------------------+
_desc|         vendor_desc|
-----+--------------------+
rate|  Curb Mobility, LLC|
rate|  Curb Mobility, LLC|
rate|  Curb Mobility, LLC|
rate|  Curb Mobility, LLC|
rate|Creative Mobile T...|
-----+--------------------+
```

Separated date and time of pickup and drop

```
+--------------------+-----------+-----------+--------------------+----------+---------+
|tpep_pickup_datetime|pickup_date|pickup_time|tpep_dropoff_datetime|drop_date |drop_time|
+--------------------+-----------+-----------+--------------------+----------+---------+
|2016-01-01 00:38:14 |2016-01-01 |00:38:14   |2016-01-01 00:52:44  |2016-01-01|00:52:44 |
|2016-01-01 00:16:05 |2016-01-01 |00:16:05   |2016-01-01 00:25:36  |2016-01-01|00:25:36 |
|2016-01-01 00:19:00 |2016-01-01 |00:19:00   |2016-01-01 00:22:41  |2016-01-01|00:22:41 |
|2016-01-01 00:42:25 |2016-01-01 |00:42:25   |2016-01-01 00:42:47  |2016-01-01|00:42:47 |
|2016-01-01 00:51:52 |2016-01-01 |00:51:52   |2016-01-01 01:02:37  |2016-01-01|01:02:37 |
+--------------------+-----------+-----------+--------------------+----------+---------+
only showing top 5 rows
```

And dropped old date_time columns of pickup and drop

```
    clean_sample_df = clean_sample_df.drop("tpep_pickup_datetime", "tpep_dropoff_datetime")
]
```

Dropped service zone and borough

```
    # If two columns with same name exist, Spark won't allow writing.
    # Drop one of each duplicate manually, based on what you want to keep.

    clean_sample_df = clean_sample_df.drop("service_zone").drop("borough")
```

Final row and columns count

```
Row count: 991530
Column count: 27
```

Final schema

```
root
 |-- DOLocationID: long (nullable = true)
 |-- PULocationID: long (nullable = true)
 |-- store_and_fwd_flag: string (nullable = true)
 |-- passenger_count: double (nullable = true)
 |-- extra: double (nullable = true)
 |-- tip_amount: double (nullable = true)
 |-- fare_amount: double (nullable = true)
 |-- total_amount: double (nullable = true)
 |-- mta_tax: double (nullable = true)
 |-- trip_distance: double (nullable = true)
 |-- VendorID: long (nullable = true)
 |-- improvement_surcharge: double (nullable = true)
 |-- payment_type: long (nullable = true)
 |-- RatecodeID: integer (nullable = true)
 |-- tolls_amount: double (nullable = true)
 |-- year: integer (nullable = true)
 |-- pickup_zone: string (nullable = true)
 |-- dropoff_zone: string (nullable = true)
 |-- tip_percentage: double (nullable = true)
 |-- distance_bucket: string (nullable = false)
 |-- payment_type_desc: string (nullable = false)
 |-- ratecode_desc: string (nullable = false)
 |-- vendor_desc: string (nullable = true)
 |-- pickup_date: date (nullable = true)
 |-- pickup_time: string (nullable = true)
 |-- drop_date: date (nullable = true)
 |-- drop_time: string (nullable = true)
```