Exercises

1. Perform cleaning on this dataset for records where passenger count is 0.

Exercises (cont.)

2. The data has several trips for which `fare_amount` is $600-700, which is implausible in the real world. These data points are considered outliers; they are highly abnormal values. There are several ways to calculate what values constitute an outlier. In this scenario, a normal fare amount is defined as between $0-100. Any amount above $100 will be considered an outlier. One way of managing outliers is replacing the aberrant values with the average value of the data points across the entire dataset.

Exercises (cont.)

3. Categorize users by *time of day* when the trip is taken. Extract the Hour of Day from the timestamp and categorize them into
    - Trips taken between 0000 hours to 1200 hours.
    - Trips taken between 1200 hours to 2400 hours.
      You can use in-built Spark functions to filter the Months and Hours and then perform a groupBy operation to get a count of users across 4 hour intervals throughout the day.

- `dayofmonth`
- `year`
- `weekofyear`
- `quarter`
- `month`
- `minute`

Exercises (cont.)

4. Use the `payment_type` column of the fares DataFrame to get a count of the usage of various payment methods. Then sort them to find out which payment methods are being used more frequently.

Exercises (cont.)

5. Extract Day of Month from Timestamp and append to DataFrame as a separate column having name `DateOfTrip`
6. Filter the `PULocationID` column to get only those trips having Location ID 132. Then perform a count operation to get the number of trips from that particular location.
7. Calculate the Average Speed for each trip by finding the duration of each trip; this can be done by calculating the difference of pickup time and the dropoff time. Then divide the distance column by the duration to get the mean speed.