**Data Preprocessing:**

1. **Import the necessary libraries:**

Pandas is imported to read the data in a data frame. Numpy is imported to create necessary arrays from the data frame, wherever required. Seaborn is imported to visualise the data. Matplotlib is also imported for visualisation.



1. **Read the data set:**

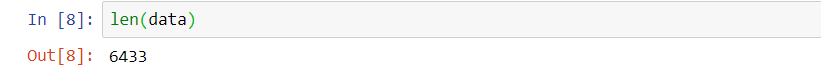
We read the data set in a Pandas data frame.



1. **Review the first 10 columns in the data set:**

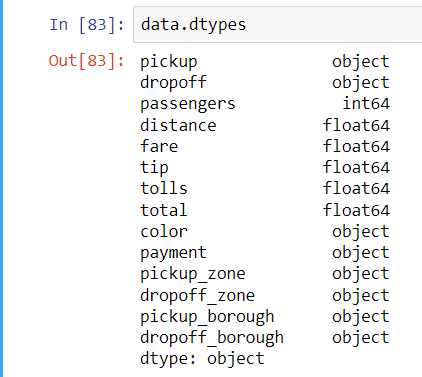


1. **Find the number of rows in the data set:**

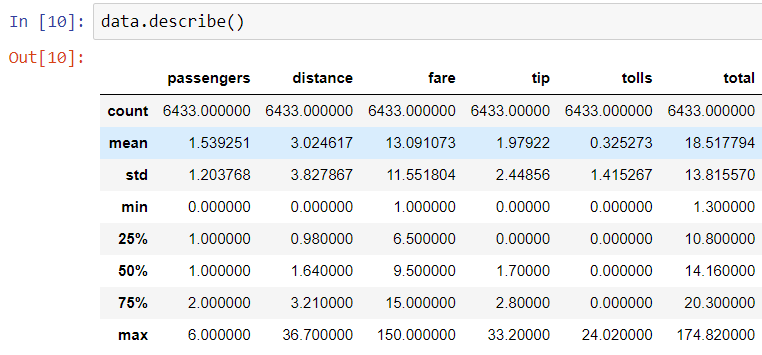


* There are 6433 rows in the data set.

1. **Find the data types of all the columns in the data set**.

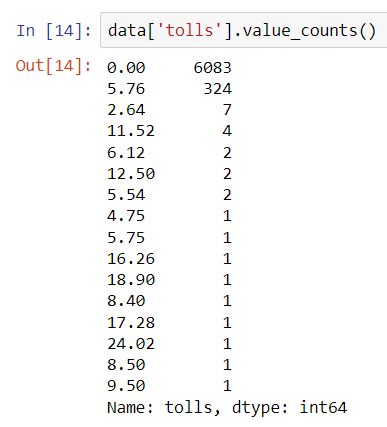


1. **Find the statistical description of all the columns in the data set:**

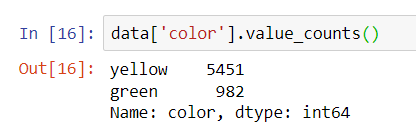


* The median value of passengers is 1. This means that for half of the trips, there is only 1 passenger.
* The median value of total fare is 14.16 and the mean value of total fare is 18.52. This means the total fare is mostly around and below these values, its not very high for most of the trips.
* For majority of trips, toll was not charged.
* For majority of trips, the tip value is very low.
* For majority of trips, the fare was less than 15 units.
* In majority of trips, a very short distance was covered.

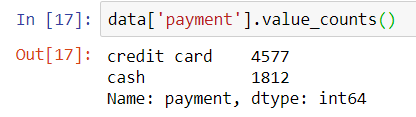
1. **Find the number of unique values for important columns the data set:**



* For 6083 trips, the toll charged was 0.

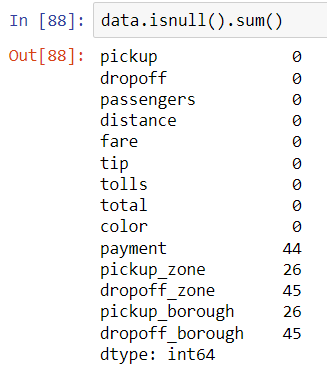


* Yellow coloured vehicle was used for majority of trips i.e. 5451 trips.

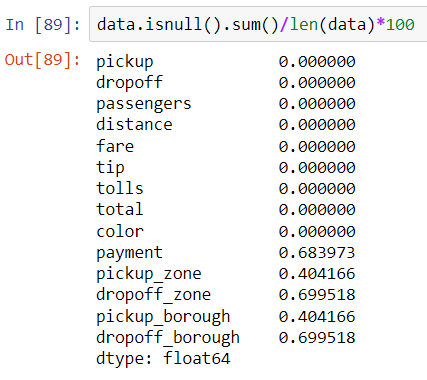


* Passengers paid via credit card for 4577 trips as opposed to 1812 trips.

1. **Find the number of null values in the data set:**

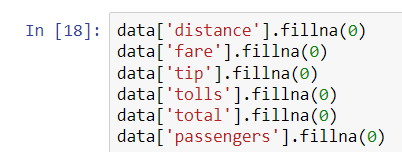


1. **Check the %age of null values in the data set**:



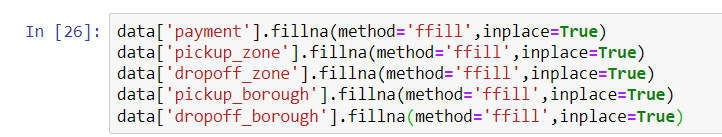
* The %age of null values is very low in this data set.

1. **Filling the Nas in numerical columns with 0:**



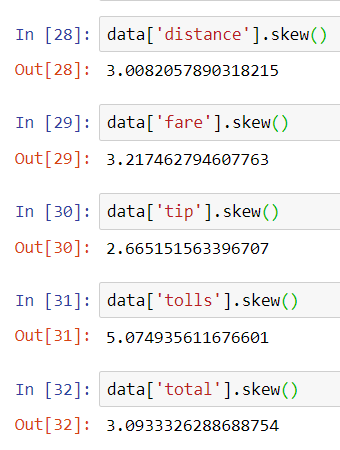
1. **For filling Nas in non-numerical columns, using forward fill:**

* In forward fill, the data in the lower numbered rows is used to fill the missing values in its consequent or very next higher row.



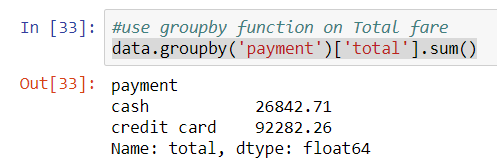
1. **Find the skewness of all the numerical columns.**

* Skewness is the degree of asymmetry observed in a probability distribution.
* Negative skew refers to longer or fatter tails on the left side of the distribution.
* Positive skew refers to longer or fatter tails on the right side of the distribution.
* These two skews refer to the direction or weight of the distribution.
* Mean of positively skewed data is greater than the median.
* Mean of negatively skewed data is less than the median.



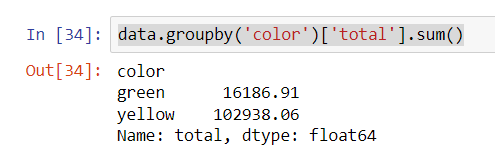
* All the above distributions for distance, fare, tip , tolls and total are positively skewed , hence the mean is greater than the median, that means most of the values are less than the mean. It can be inferred that total bill paid by majority of passengers was less than the average bill.

1. **Sum of total by payment, using group by:**



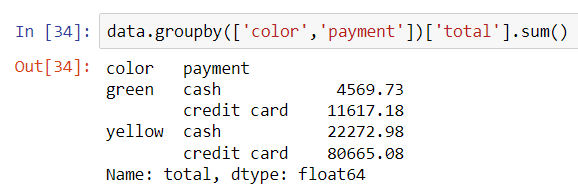
* Majority of payments were done through credit card.

1. **Sum of total by colour, using group by:**



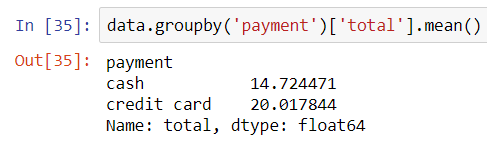
* Majority of the trips were completed in yellow-coloured vehicles.

1. **Sum of total by color and payment, using group by:**



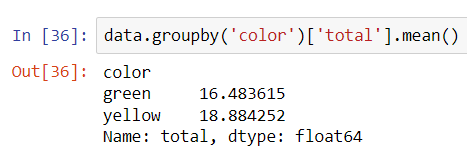
* Number of credit card payments was highest for green-colored vehicles

1. **Mean of total, group by payment**



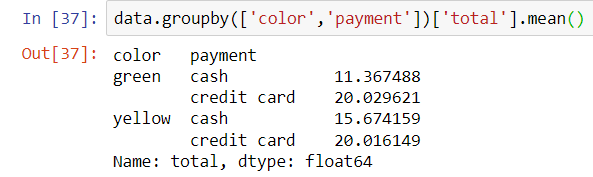
* There is not much difference in the total cash, irrespective of payment mode.

1. **Mean of total, group by color:**

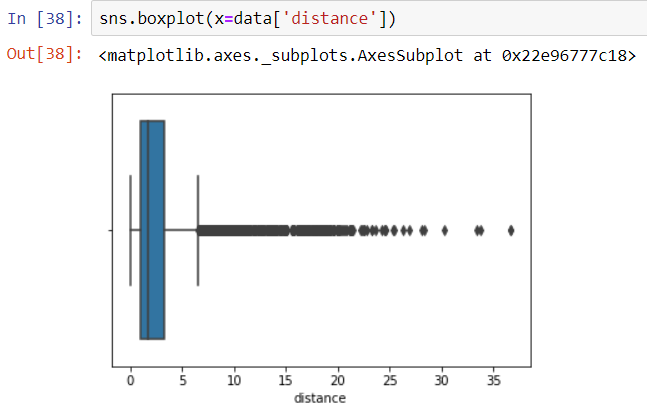


* Total payment was almost equal for both yellow as well as green coloured vehicles.

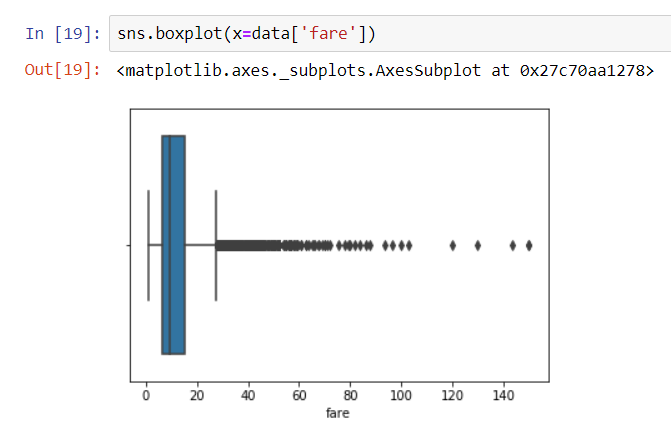
1. **Mean of total, group by payment and color:**



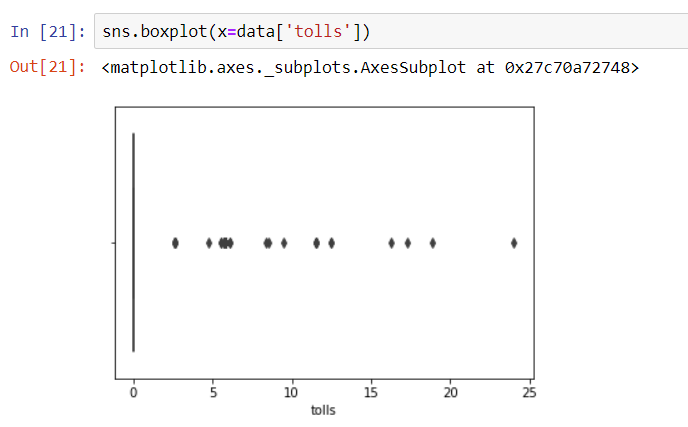
1. **Detecting outliers through boxplot**

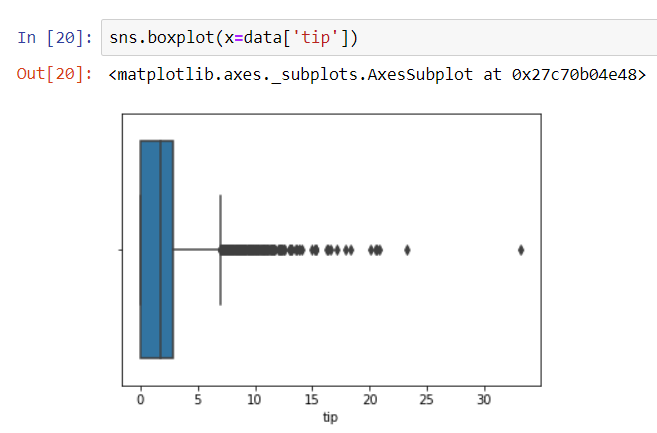


* There are many outliers in distance parameter.

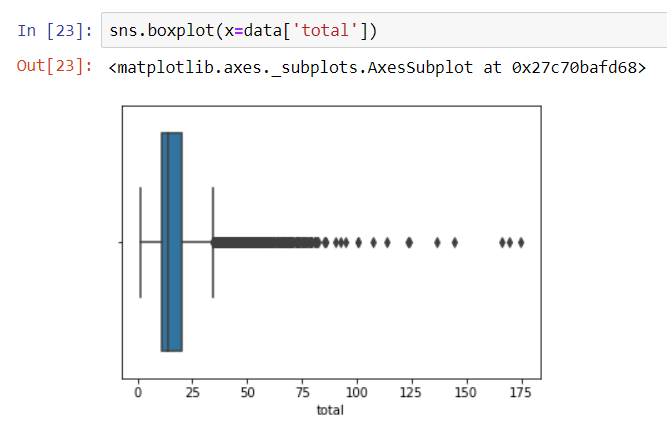


* There are few outliers in fare parameter.



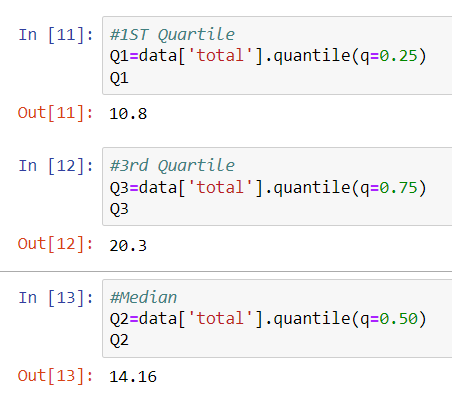


* There are few outliers in tip parameter.

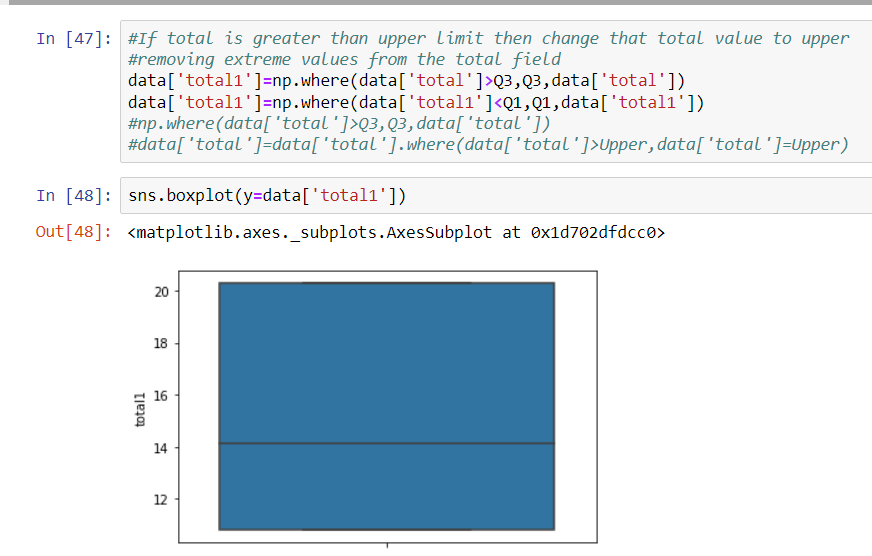


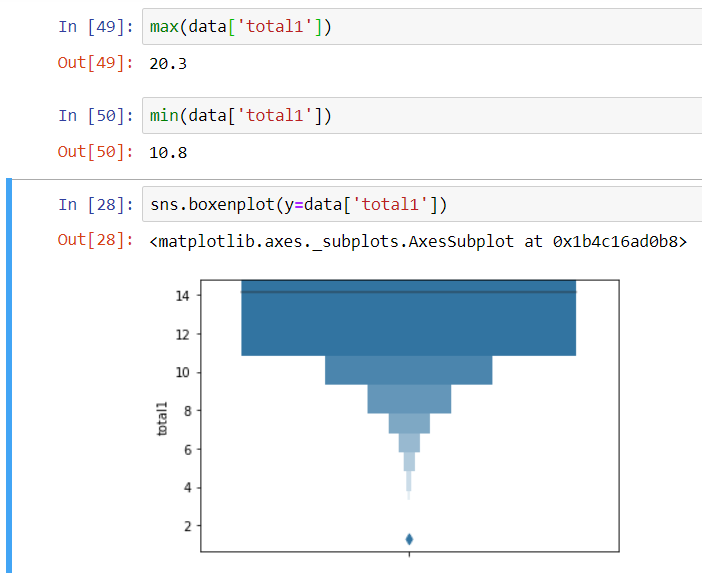
* There are few outliers in total parameter field.

1. **Finding the 1st quartile, 3rd quartile and median for total fare:**

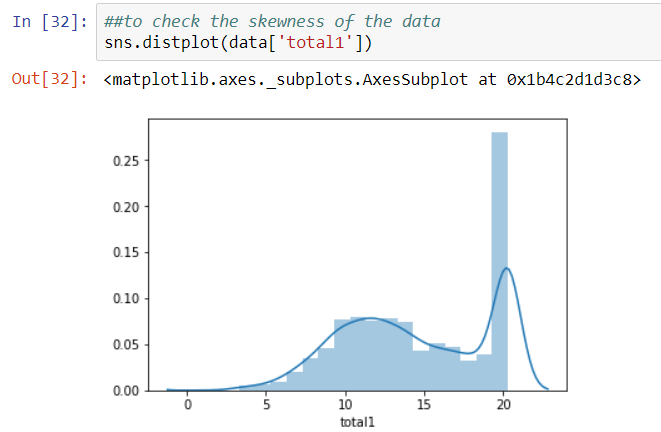


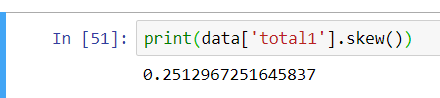
1. **Removing all outliers above 3rd quartile and below 1st quartile. Assigning the value of 1st quartile to all the outliers below it and the value of 3rd quartile to all the outliers above it. Doing this calculation only for Total column:**





1. **Checking the skewness of data:**

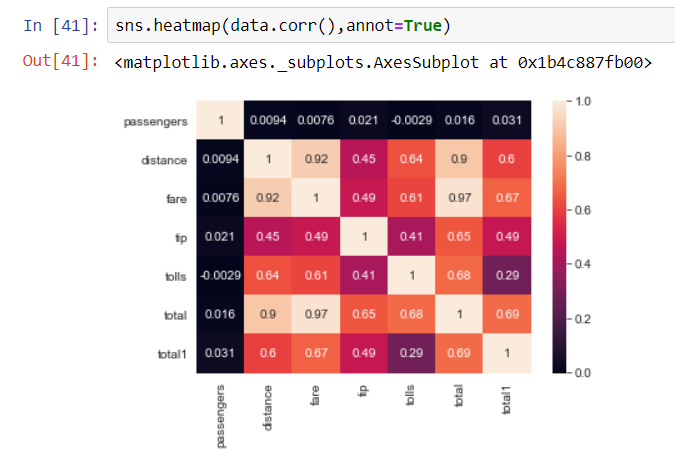




* The total1, i.e., the total value adjusted after removing outliers, is slightly positively skewed towards right.

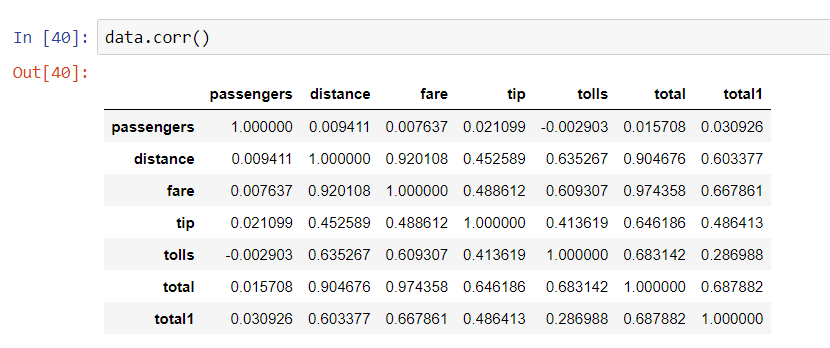
**Exploratory Data Analysis:**

1. **Finding the correlation among numerical variables**



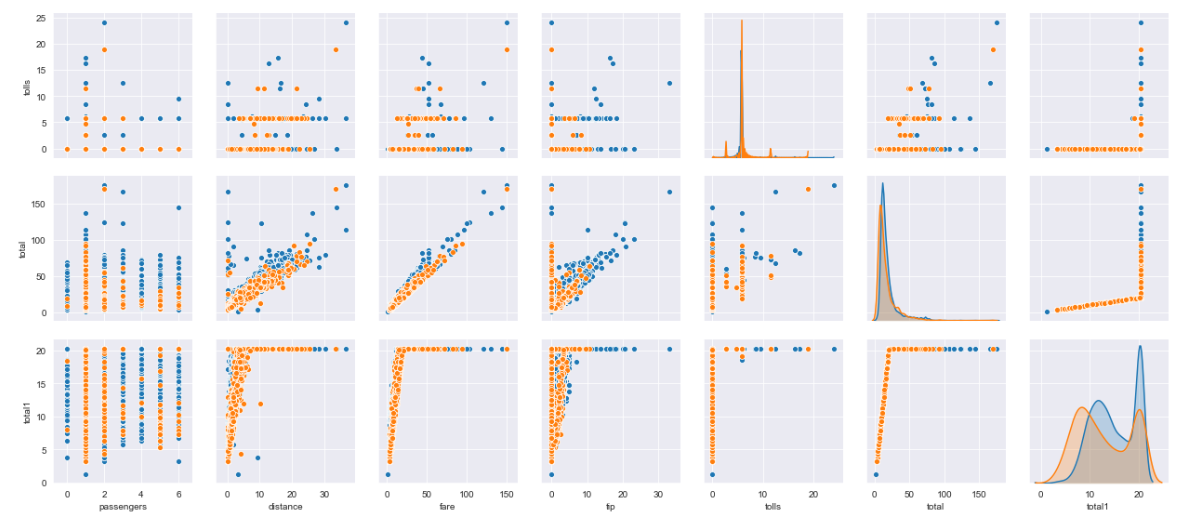
correlation between the following entities is more than 0.5:

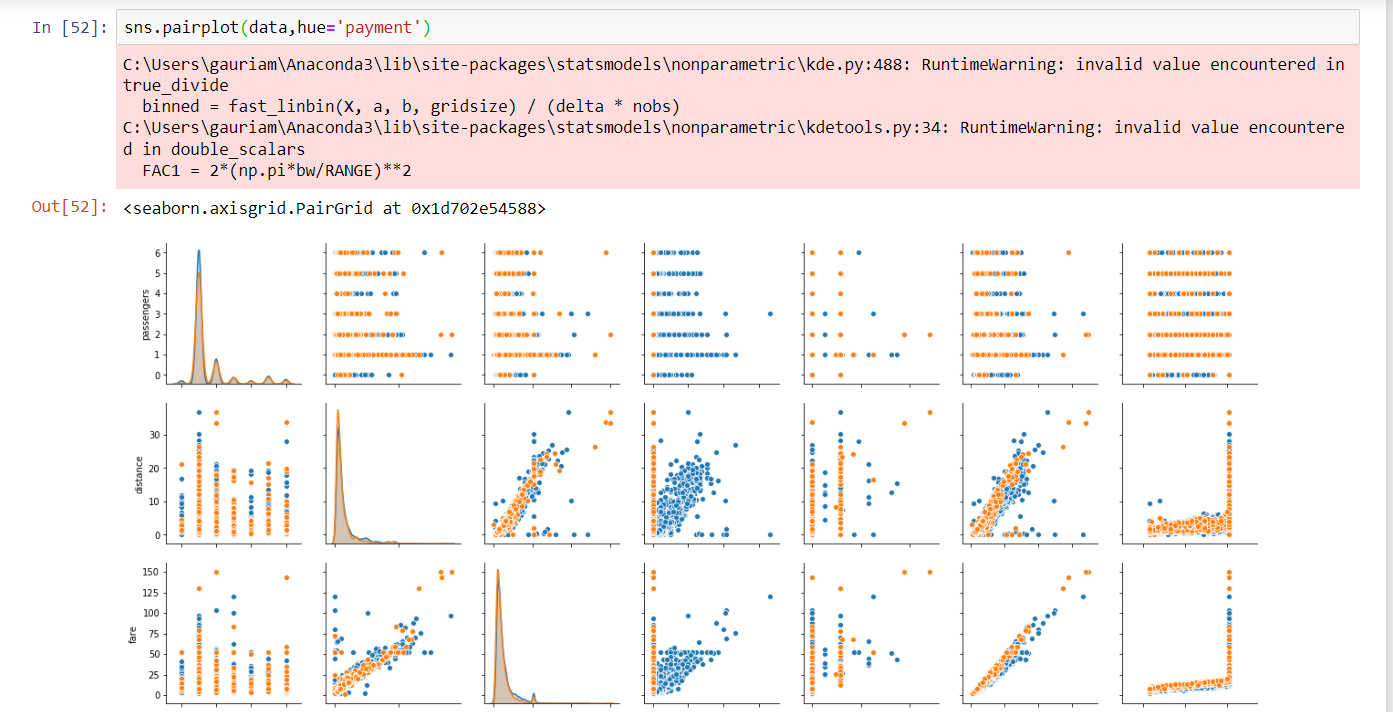
* Distance and Fare
* Distance and Bills
* Distance and Total1 (Total variable adjusted after removing outliers above the 1st and 3rd quartiles)
* Fare and Bills
* Fare and Total
* Tip and Total
* Tolls and Distance
* Tolls and Fare
* Tolls and Total
* Total1 and Fare



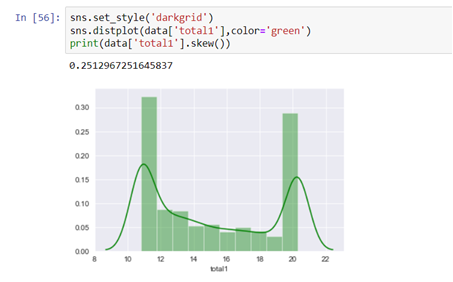
1. **Pairplot:**

****

****

****

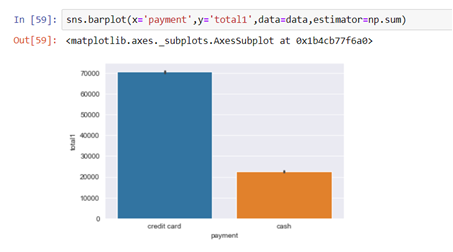
1. **Distribution Plot of Total1 variable:**

****

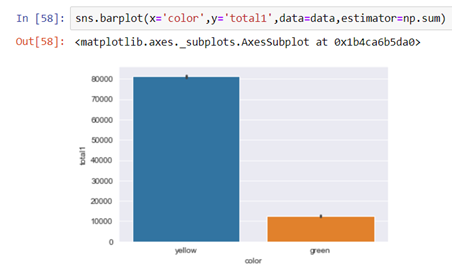
* Majority of total fare is in the range 10-12 and around 20.

1. **Bar charts**

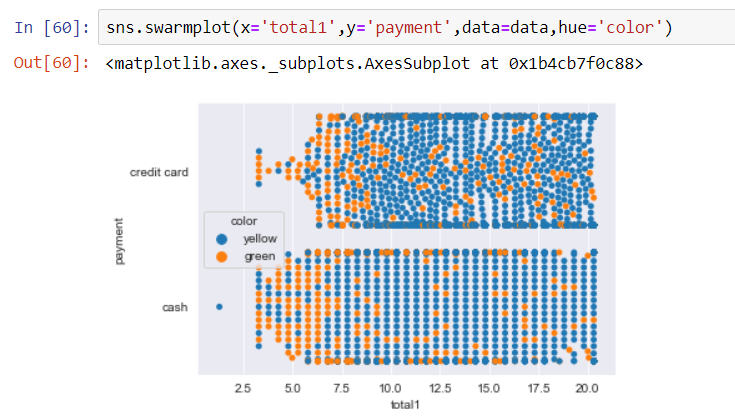
* Barplot of Payment mode v/s total1 (Total fare adjusted after removing outliers beyond the 1st and 3rd quartile respectively). We can see that 70,000 worth of total fare was paid through credit card v/s 20,000 worth of total fare through cash.

****

* Barplot of color v/s total1. We can see that 80,000 worth of total fare was spent by passengers in blue car as opposed to 10,000 worth of total fare for passengers in green car.

****

1. **Swarm plot:**



1. **Line charts**

