

Data Analysis

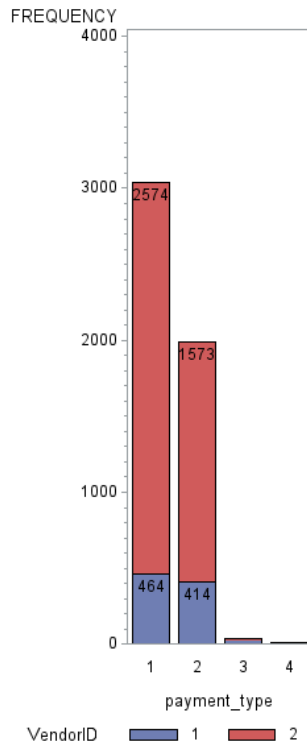
For our data analysis we used a dataset on New York taxi rides of two different taxi companies.

Variable	N	Mean	Std Dev	Minimum	Maximum
passenger_count	5064	1.35	1.04	0.00	6.00
trip_distance	5064	3.61	4.34	0.00	119.82
fare_amount	5064	14.61	12.47	-18.00	105.00
tip_amount	5064	1.10	2.13	0.00	40.00
tolls_amount	5064	0.25	1.17	0.00	11.52
total_amount	5064	16.99	13.69	-18.00	129.15

In the next step of our script we got rid of the negative values in fare_amount and total_amount as those values do not make sense.

Covariance Table				
	PAYMENT_TYPE	FARE_AMOUNT	TRIP_DISTANCE	PASSENGER_COUNT
PAYMENT_TYPE	0.261	-1.798	-0.504	-0.009
FARE_AMOUNT	-1.798	155.0296	46.886	0.100
TRIP_DISTANCE	-0.504	46.886	18.848	0.114
PASSENGER_COUNT	-0.009	0.100	0.114	1.071

Frequency of vendor and passenger amount



Scatterplot of tip_amount and fare_amount.

