Week 6 File ingestion and schema validation

Abida Bhatti

Batch Code: LISUM01

July 17,2021

submitted to Github

Data Source
https://www.kaggle.com/new-york-city/nyc-parking-tickets?select=Parking_Violations_Issued_-_Fiscal_Year_2016.csvData file name
Parking_Violations_Issued_-_Fiscal_Year_2016.csv

I have completed following steps for assignment 6

1. Used csv/text file of 2G
2. I used pandas data frame with the options of chunksize= 100,000.
    a) Process each chunk one by one
    b) Removed columns that are not being used for my analysis
    c) We used only 6 columns out of 51 columns
    d) Remove data where we don't have valid state
    e) Update default value of "0" where violation location is empty
3. Appended final data final data frame
4. Created YAML file for schema validation
5. Added following functions to Common function utility file:
    a) Functions that return the basic statistic of the file
    b) Functions to convert the file to gz format
    c) Function to convert common to pipe sign(|)
6. Update Read me file in Github

Screen shot of basic summary statistics of the data frames

['vehicle_expiration date', 'unregistered_vehicle?']
Columns name and column length validation failed
Following File columns are not in the YAML file ['vehicle_expiration_date', 'unregistered_vehicle']
Following File columns are not in the following uploaded file ['vehicle_expiration date', 'unregistered_vehicle?']
Total Number of Columns in the raw data set: 51
Total Number of Columns in the data set: 6
Total Number of Rows in the raw data set: 10555080
File Size is : 2101501.765625 KB

Raw file size = 2.052 GB

Raw data Sample

Summons Nu	Plate ID	Registrati	Plate Type	Issue Date	Violation	Vehicle Bo	Vehicle M	Issuing Ag	Street Coc	Street Coc	Street Coc	Vehicle Ex	Violation	Violation	Issuer Pre	Issuer Coc	Issuer Cor	Issuer Squ	Violation	Time First	Violation	Violation	House Nu	Street Nar	Intersecti	Date First	Law Sectio	Sub
4608806129	ZUY52G	NJ	PAS	7/23/2015	36	WAGO	CHEVR	V	0	0	0	0		0	0	0			0906A					NB BAY PK	T	0	1180	B
4608806142	GTC5999	NY	PAS	7/23/2015	36	SUBN	ACURA	V	0	0	0	0		0	0	0			0908A					SB BAYCH	CRAWFOR	0	1180	B
4608806154	JFV9493	PA	PAS	7/23/2015	36	SDN	LINCO	V	0	0	0	0		0	0	0			0908A					NB KINGS	H ST	0	1180	B
4608806178	GUY3696	NY	PAS	7/23/2015	36	SUBN	TOYOT	V	0	0	0	0		0	0	0			0908A					SB BAYCH	CRAWFOR	0	1180	B
4608806221	T19DWV	NJ	PAS	7/23/2015	36	WAGO	ME/BE	V	0	0	0	0		0	0	0			0909A					NB BAY PK	T	0	1180	B
4608806245	GDY4545	NY	PAS	7/23/2015	36	SUBN	HONDA	V	0	0	0	0		0	0	0			0909A					EB BEACH	@ BEACH	0	1180	B
4608806269	GTA3269	NY	PAS	7/23/2015	36	SUBN	INFIN	V	0	0	0	0		0	0	0			0909A					SB WOOD	FURMANV	0	1180	B
4608806282	GPS3941	NY	PAS	7/23/2015	36	SUBN	GMC	V	0	0	0	0		0	0	0			0910A					WB GOETH	ULES DR	0	1180	B
4608806294	GNF2093	NY	PAS	7/23/2015	36	SUBN	DODGE	V	0	0	0	0		0	0	0			0910A					EB BEACH	@ BEACH	0	1180	B
4608806336	GRV9280	NY	PAS	7/23/2015	36	SUBN	JEEP	V	0	0	0	0		0	0	0			0911A					SB WOOD	FURMANV	0	1180	B
4608806348	GWH8714	NY	PAS	7/23/2015	36	4DSD	TOYOT	V	0	0	0	0		0	0	0			0911A					SB WOOD	FURMANV	0	1180	B
4608806373	XBHE87	NJ	PAS	7/23/2015	36	VAN	CHEVR	V	0	0	0	0		0	0	0			0912A					SB BAYCH	CRAWFOR	0	1180	B
4608806385	GND6000	NY	PAS	7/23/2015	36	4DSD	HYUND	V	0	0	0	0		0	0	0			0912A					EB FLATLA	LTON ST	0	1180	B
4608806403	GPV6020	NY	PAS	7/23/2015	36	SUBN	INFIN	V	0	0	0	0		0	0	0			0913A					EB HILLSID	RSONS BL	0	1180	B
4608806439	GNT1157	NY	OMS	7/23/2015	36	SUBN	NISSA	V	0	0	0	0		0	0	0			0913A					SB WOOD	FURMANV	0	1180	B
4608806452	GSU6660	NY	PAS	7/23/2015	36	SUBN	ACURA	V	0	0	0	0		0	0	0			0913A					EB E GUN	ARNES AV	0	1180	B
4608806464	JVA1731	PA	PAS	7/23/2015	36	SDN	CHEVR	V	0	0	0	0		0	0	0			0913A					SB KINGS	N AVE	0	1180	B
4608806476	GCJ7464	NY	PAS	7/23/2015	36	4DSD	HYUND	V	0	0	0	0		0	0	0			0914A					EB BEACH	@ BEACH	0	1180	B
4608806488	GTB5341	NY	PAS	7/23/2015	36	4DSD	BMW	V	0	0	0	0		0	0	0			0915A					WB HILLSI	6TH ST	0	1180	B
4608806506	GFS6123	NY	PAS	7/23/2015	36	4DSD	NISSA	V	0	0	0	0		0	0	0			0915A					SB BAYCH	CRAWFOR	0	1180	B
4608806543	GPK1247	NY	PAS	7/23/2015	36	SUBN	CHRYS	V	0	0	0	0		0	0	0			0916A					SB BAYCH	CRAWFOR	0	1180	B
4608806555	GPU1087	NY	PAS	7/23/2015	36	4DSD	NISSA	V	0	0	0	0		0	0	0			0917A					EB ROCKA	00TH ST	0	1180	B
4608806622	J71DFP	NJ	PAS	7/23/2015	36	4 DR	HONDA	V	0	0	0	0		0	0	0			0918A					WB QUEEI	H ST	0	1180	B
4608806634	L98CSD	NJ	PAS	7/23/2015	36	4 DR	HONDA	V	0	0	0	0		0	0	0			0918A					WB GOETH	ULES DR	0	1180	B
4608806695	GHN2582	NY	PAS	7/23/2015	36	SUBN	CHRYS	V	0	0	0	0		0	0	0			0920A					EB SEAVIE	00TH ST	0	1180	B
4608806701	GTJ8740	NY	PAS	7/23/2015	36	4DSD	BMW	V	0	0	0	0		0	0	0			0920A					WB JAMA	TH ST	0	1180	B
4608806725	T661697C	NY	OMT	7/23/2015	36	4DSD	LINCO	V	0	0	0	0		0	0	0			0920A					EB HILLSID	RSONS BL	0	1180	B
4608806737	GRD1793	NY	PAS	7/23/2015	36	4DSD	TOYOT	V	0	0	0	0		0	0	0			0920A					EB HILLSID	RSONS BL	0	1180	B
4608806774	N91FKN	NJ	PAS	7/23/2015	36	WAGO	JEEP	V	0	0	0	0		0	0	0			0921A					WB QUEEI	H ST	0	1180	B
4608806786	GGC8019	NY	PAS	7/23/2015	36	4DSD	CHRYS	V	0	0	0	0		0	0	0			0921A					EB SEAVIE	00TH ST	0	1180	B
4608806798	GXC1876	NY	PAS	7/23/2015	36	SUBN	INFIN	V	0	0	0	0		0	0	0			0921A					EB ROCKA	00TH ST	0	1180	B
4608806830	JLC4784	PA	PAS	7/23/2015	36	SW	BMW	V	0	0	0	0		0	0	0			0922A					NB MERRI	NDEN BLV	0	1180	B
4608806890	GJZ4500	NY	PAS	7/23/2015	36	4DSD	NISSA	V	0	0	0	0		0	0	0			0923A					SB KINGS	N AVE	0	1180	B
4608806919	JMN5062	PA	PAS	7/23/2015	36	SW	HONDA	V	0	0	0	0		0	0	0			0923A					EB FLATLA	LTON ST	0	1180	B
4608806920	GVV7194	NY	PAS	7/23/2015	36	SUBN	VOLKS	V	0	0	0	0		0	0	0			0923A					EB 65TH S	E	0	1180	B
4608806932	T618695C	NY	OMT	7/23/2015	36	SUBN	TOYOT	V	0	0	0	0		0	0	0			0924A					EB 65TH S	E	0	1180	B

Final data set Sample