Used car Data Intake Report

Total number of observations	750010
Total number of files	1
Total number of features	60
Base format of the file	CSV
Size of the data	2.1 GB

```
# Read config file
import testutility as util
config_data = util.read_config_file("file.yaml")
```

```
#inspecting data of config file
config_data
```

```
{'file_type': 'csv',
 'bucket': 'used_car',
 'fold_name': 'data',
 'file_name': 'used_cars_data_sample',
 'inbound_delimiter': ',',
 'outbound_delimiter': '|',
 'skip_leading_rows': 1,
 'columns': ['vin', 'bed', 'year']}
```

```
#inspecting data of config file
config_data['file_name']
```

```
'used_cars_data_sample'
```

Launcher | Untitled.ipynb | Code | git | Python 3

```
[1]: import pandas as pd
```

```
[2]: used_car= pd.read_csv('gs://used_car/data/used_cars_data.csv')
```

```
/tmp/ipykernel_10908/3197685807.py:1: DtypeWarning: Columns (11) have mixed types. Specify dtype option on import or set low_memory=False.
  used_car= pd.read_csv('gs://used_car/data/used_cars_data.csv')
```

```
[3]: used_car.shape
```

```
[3]: (3000040, 66)
```

```
[4]: 3000040/4
```

```
[4]: 750010.0
```

```
[5]: used_car.head()
```

[5]:

:abin	city	city_fuel_economy	combine_fuel_economy	...	transmission	transmission_display	trimId	trim_name	vehicle_damage_category	wheel_syst
NaN	Bayamon	NaN	NaN	...	A	9-Speed Automatic Overdrive	t83804	Latitude FWD	NaN	F
NaN	San Juan	NaN	NaN	...	A	9-Speed Automatic Overdrive	t86759	S AWD	NaN	A
NaN	Guaynabo	17.0	NaN	...	M	6-Speed Manual	t58994	Base	NaN	A
NaN	San Juan	NaN	NaN	...	A	8-Speed Automatic Overdrive	t86074	V6 HSE AWD	NaN	A

```
#validate the header of the file
util.col_header_val(df,config_data)
```

```
column name and column length validation failed
Following File columns are not in the YAML file ['engine_cylinders', 'has_accidents', 'owner_count', 'bed_length', 'fuel_tank_volume', 'cit
y', 'main_picture_url', 'longitude', 'seller_rating', 'theft_title', 'transmission_display', 'trimid', 'combine_fuel_economy', 'horsepower',
'major_options', 'body_type', 'width', 'latitude', 'listed_date', 'is_new', 'is_oemcpo', 'fuel_type', 'description', 'franchise_dealer', 'ma
ke_name', 'unnamed_0', 'listing_color', 'salvage', 'listing_id', 'height', 'price', 'savings_amount', 'transmission', 'interior_color', 'cit
y_fuel_economy', 'model_name', 'frame_damaged', 'front_legroom', 'engine_type', 'wheelbase', 'cabin', 'bed_height', 'daysonmarket', 'is_cert
ified', 'maximum_seating', 'highway_fuel_economy', 'trim_name', 'sp_name', 'fleet', 'iscab', 'sp_id', 'length', 'exterior_color', 'mileage',
'power', 'franchise_make', 'vehicle_damage_category', 'engine_displacement', 'torque', 'wheel_system', 'back_legroom', 'wheel_system_displa
y', 'is_cpo', 'dealer_zip']
Following YAML columns are not in the file uploaded []
0
```

[22]:
```
if util.col_header_val(df,config_data)==0:
    print("validation failed")
    # write code to reject the file
else:
    print("col validation passed")
    # write the code to perform further action
    # in the pipleine
```

```
column name and column length validation failed
Following File columns are not in the YAML file ['engine_cylinders', 'has_accidents', 'owner_count', 'bed_length', 'fuel_tank_volume', 'cit
y', 'main_picture_url', 'longitude', 'seller_rating', 'theft_title', 'transmission_display', 'trimid', 'combine_fuel_economy', 'horsepower',
'major_options', 'body_type', 'width', 'latitude', 'listed_date', 'is_new', 'is_oemcpo', 'fuel_type', 'description', 'franchise_dealer', 'ma
ke_name', 'unnamed_0', 'listing_color', 'salvage', 'listing_id', 'height', 'price', 'savings_amount', 'transmission', 'interior_color', 'cit
y_fuel_economy', 'model_name', 'frame_damaged', 'front_legroom', 'engine_type', 'wheelbase', 'cabin', 'bed_height', 'daysonmarket', 'is_cert
ified', 'maximum_seating', 'highway_fuel_economy', 'trim_name', 'sp_name', 'fleet', 'iscab', 'sp_id', 'length', 'exterior_color', 'mileage',
'power', 'franchise_make', 'vehicle_damage_category', 'engine_displacement', 'torque', 'wheel_system', 'back_legroom', 'wheel_system_displa
y', 'is_cpo', 'dealer_zip']
Following YAML columns are not in the file uploaded []
validation failed
```

```
print("columns of files are:" ,df.columns)
print("columns of YAML are:" ,config_data['columns'])
```

```
columns of files are: Index(['unnamed_0', 'vin', 'back_legroom', 'bed', 'bed_height', 'bed_length',
       'body_type', 'cabin', 'city', 'city_fuel_economy',
       'combine_fuel_economy', 'daysonmarket', 'dealer_zip', 'description',
       'engine_cylinders', 'engine_displacement', 'engine_type',
       'exterior_color', 'fleet', 'frame_damaged', 'franchise_dealer',
       'franchise_make', 'front_legroom', 'fuel_tank_volume', 'fuel_type',
       'has_accidents', 'height', 'highway_fuel_economy', 'horsepower',
       'interior_color', 'iscab', 'is_certified', 'is_cpo', 'is_new',
       'is_oemcpo', 'latitude', 'length', 'listed_date', 'listing_color',
       'listing_id', 'longitude', 'main_picture_url', 'major_options',
       'make_name', 'maximum_seating', 'mileage', 'model_name', 'owner_count',
       'power', 'price', 'salvage', 'savings_amount', 'seller_rating', 'sp_id',
       'sp_name', 'theft_title', 'torque', 'transmission',
       'transmission_display', 'trimid', 'trim_name',
       'vehicle_damage_category', 'wheel_system', 'wheel_system_display',
       'wheelbase', 'width', 'year'],
      dtype='object')
columns of YAML are: ['vin', 'bed', 'year']
```

```
df.to_csv('gs://used_car/data/used_car_sample.txt.gz', sep='|', compression='gzip', index=False)
```