In [1]:
```
import pandas as pd
import pandas.io.sql as psql
import re
import sqlite3
import psycopg2
from pandasql import sqldf
from matplotlib import pyplot as plt
from statsmodels.formula.api import ols
from sqlalchemy import create_engine
import pandas.io.sql as psql

con = psycopg2.connect('postgresql://aqlxfqja:F6bE-fv-jhA_VaaLV284XVgxXOLNAp_2@rosie

engine = create_engine('postgresql://aqlxfqja:F6bE-fv-jhA_VaaLV284XVgxXOLNAp_2@rosie
plt.rcParams['figure.figsize'] = [10, 5]
```

In [2]:
```
data = pd.read_csv('vehicles.csv')
```

In [3]:
```
data = data.drop(columns = ['url','region_url', 'image_url', 'VIN', 'description','c
data['id'] = data['id'].astype("str")
data['posting_date'] = data['posting_date'].astype("str")
```

In [4]:
```
posting_year = []
for x in data['posting_date']:
    if len(re.findall('[\d,]+',x)) == 0:
        posting_year.append(None)
    else:
        posting_year.append(re.findall('[\d,]+',x)[0])

data['posting_year'] = posting_year
data['posting_year'] = data['posting_year'].astype("float64")
```

In [5]:
```
data['years_old'] = data['posting_year'] - data['year']
```

In [1]:
```
data.to_csv('cleaned_vehicles.csv',index = False)
pd.read_csv("cleaned_vehicles.csv").to_sql('us_carsales_full', engine, index = False
```

In [ ]:
```
data.info()
```

In [ ]:
```
%load_ext sql

# Make a copy of dataset
%%sql postgresql://aqlxfqja:F6bE-fv-jhA_VaaLV284XVgxXOLNAp_2@rosie.db.elephantsql.co
CREATE TABLE IF NOT EXISTS us_carsales_v1 AS
(SELECT *
  FROM us_carsales_full)

# Delete record where manufacturer/model/year is null
%%sql postgresql://aqlxfqja:F6bE-fv-jhA_VaaLV284XVgxXOLNAp_2@rosie.db.elephantsql.co
DELETE
  FROM us_carsales_v1
WHERE manufacturer IS NULL
    OR model IS NULL
    OR year IS NULL
```

In [ ]:
```
print("Data clean is done")
```