Cars Dataset Visualization

Auto-mpg dataset

- Dataset with 398 cars produced between 1970 and 1982.
- We start with a short exploratory data analysis.

Column	Description
mpg	Miles/(US) gallon
cylinders	Number of cylinders
displacement	Displacement (cu.in.)
horsepower	Gross horsepower
weight	Weight (lbs)
acceleration	Time to go 0-60 mph in seconds
model_year	Model year
origin	Region of origin
name	Name of model

Description of numeric columns

	mpg	cylinders	displacement	horsepower	weight	acceleration	model_year
count	398.000000	398.000000	398.000000	392.000000	398.000000	398.000000	398.000000
mean	23.514573	5.454774	193.425879	104.469388	2970.424623	15.568090	76.010050
std	7.815984	1.701004	104.269838	38.491160	846.841774	2.757689	3.697627
min	9.000000	3.000000	68.000000	46.000000	1613.000000	8.000000	70.000000
25%	17.500000	4.000000	104.250000	75.000000	2223.750000	13.825000	73.000000
50%	23.000000	4.000000	148.500000	93.500000	2803.500000	15.500000	76.000000
75%	29.000000	8.000000	262.000000	126.000000	3608.000000	17.175000	79.000000
max	46.600000	8.000000	455.000000	230.000000	5140.000000	24.800000	82.000000

Using .describe() we get some information about the columns with numeric values.

We note that the following three columns has a smaller number of unique values:

```
cars["model_year"].value_counts(sort=False)
```

✓ 0.4s

```
70    29
71    28
72    28
73    40
74    27
75    30
76    34
77    28
78    36
79    29
80    29
81    29
82    31
Name: model_year, dtype: int64
```

```
cars["origin"].value_counts()
```

✓ 0.3s

```
usa       249
japan      79
europe     70
Name: origin, dtype: int64
```

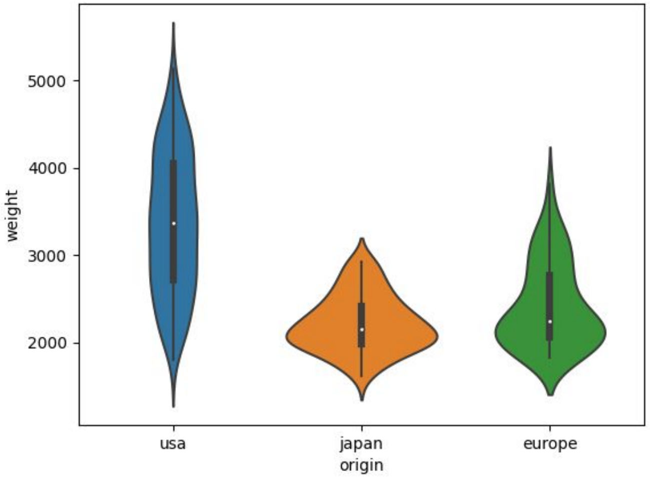
```
cars["cylinders"].value_counts()
```

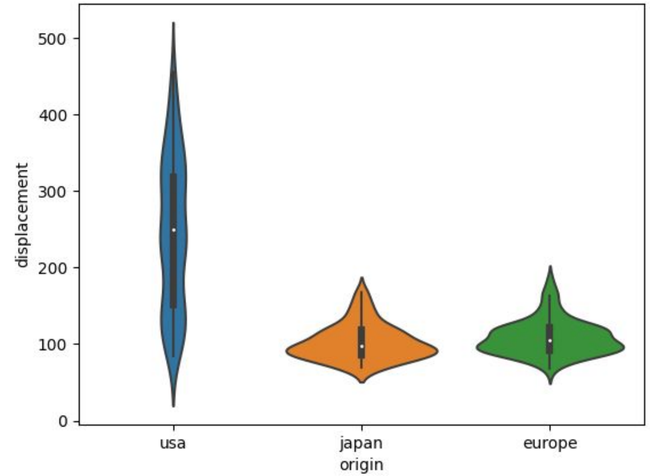
✓ 0.5s

```
4    204
8    103
6     84
3      4
5      3
Name: cylinders, dtype: int64
```

Violinplots grouped by origin

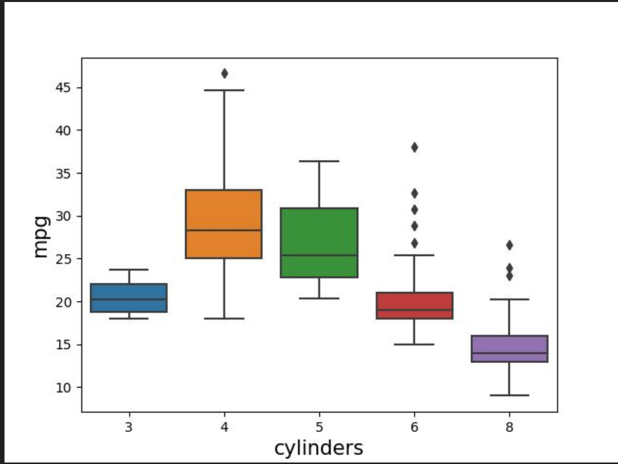
We note that the three origins vary considerably according to weight and displacement. Where the U.S.-made cars are both heavier and has larger engines than the cars made in Europe and Japan.

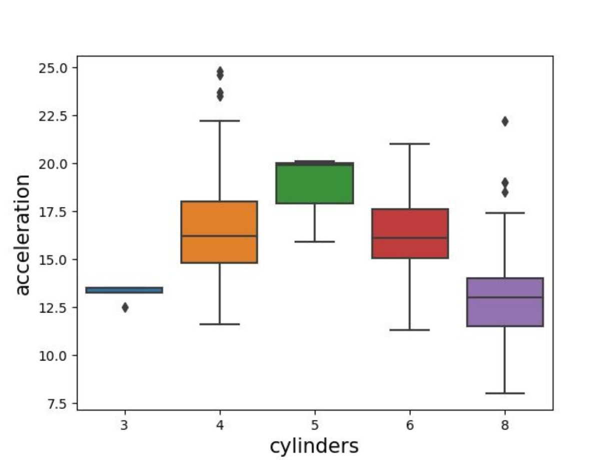


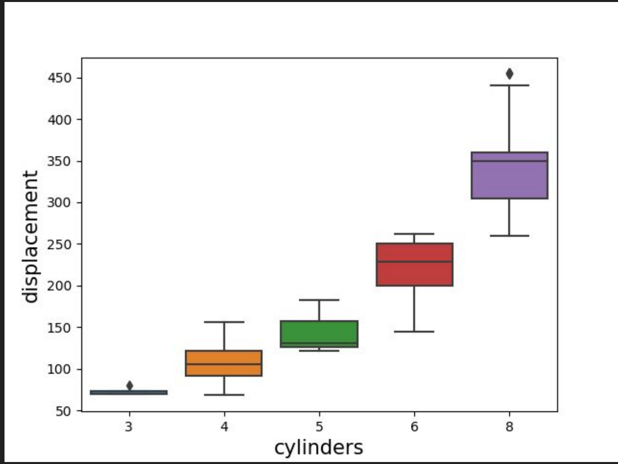


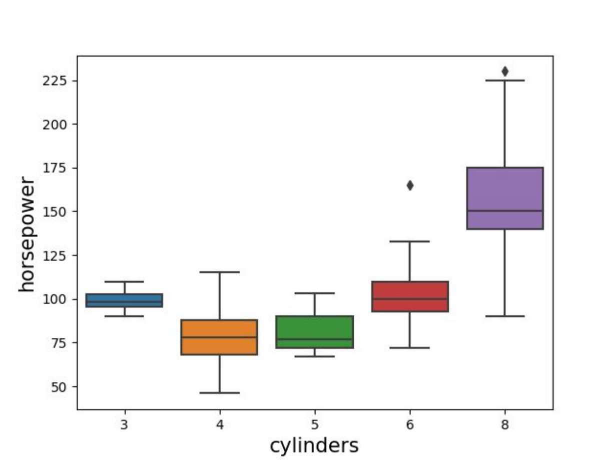
Boxplots grouped by number of cylinders

When we group by cylinders we note a few different patterns. These might be interesting to analyze further, we must note though that we only have seven observations for 3 and 5 cylinders combined.



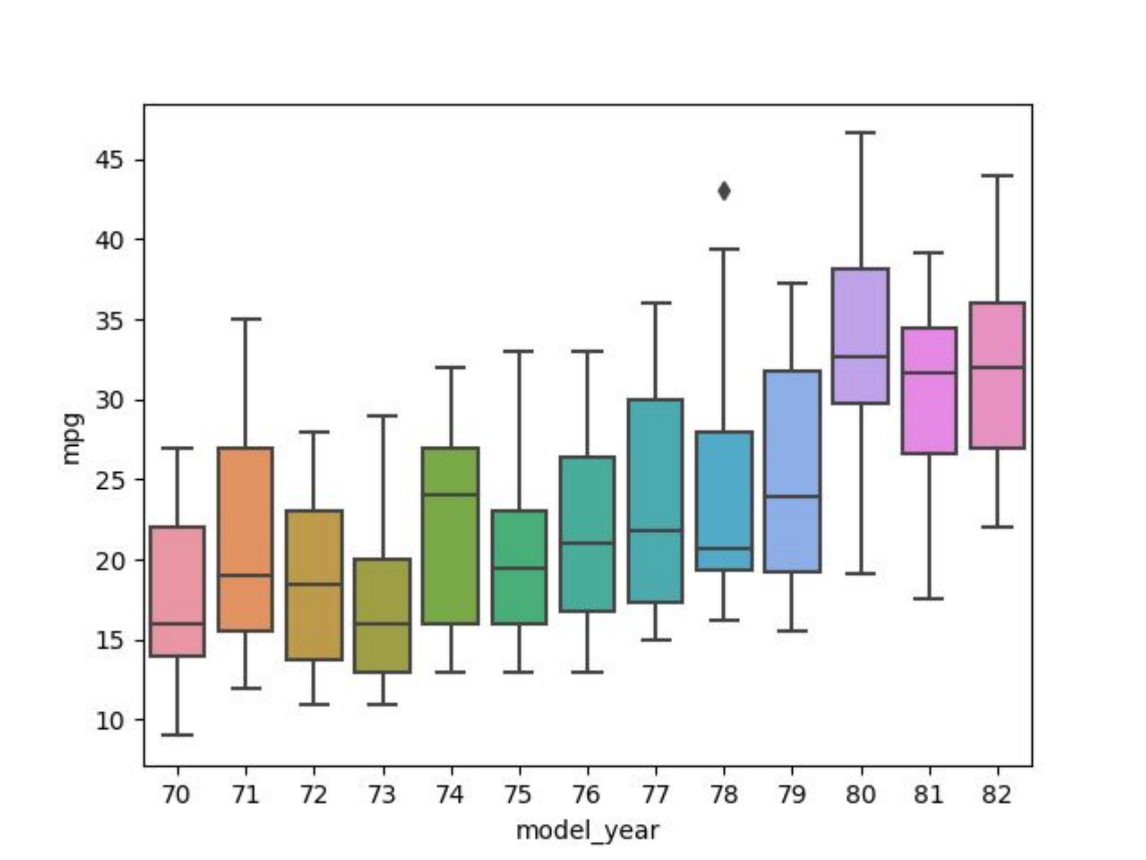






Miles per gallon

Grouping by year and analyzing the change in fuel consumption we see a possible increase in mpg over time that we can analyze further.



Questions

- Can we be confident that the mileage for our population has increased over time?
- Is there another variable that also changes with time that we can control for in our analysis?

Is mpg correlated with model_year?

- Yes! With a very low p-value ~10e-36 and with coefficient 1.23.
- We must note that other variables have an even stronger correlation with mpg, for example horsepower.
- We also note that model_year is correlated with horsepower.

```
results = smf.ols('mpg ~ model_year', data=cars).fit()
results.summary()
```

✓ 0.6s

OLS Regression Results

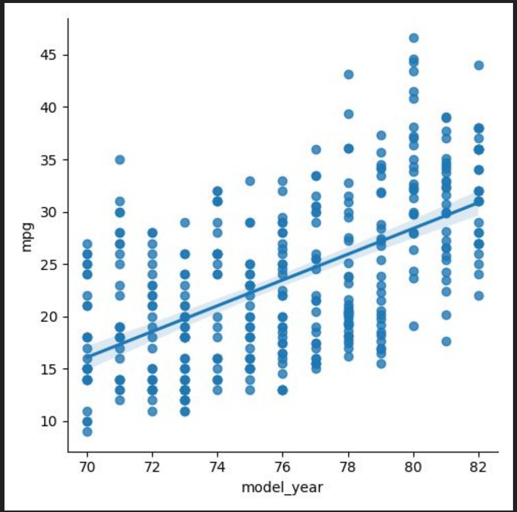
Dep. Variable:	mpg	R-squared:	0.337
Model:	OLS	Adj. R-squared:	0.335
Method:	Least Squares	F-statistic:	198.3
Date:	Tue, 31 Jan 2023	Prob (F-statistic):	1.08e-36
Time:	13:32:04	Log-Likelihood:	-1280.6
No. Observations:	392	AIC:	2565.
Df Residuals:	390	BIC:	2573.
Df Model:	1		
Covariance Type:	nonrobust		

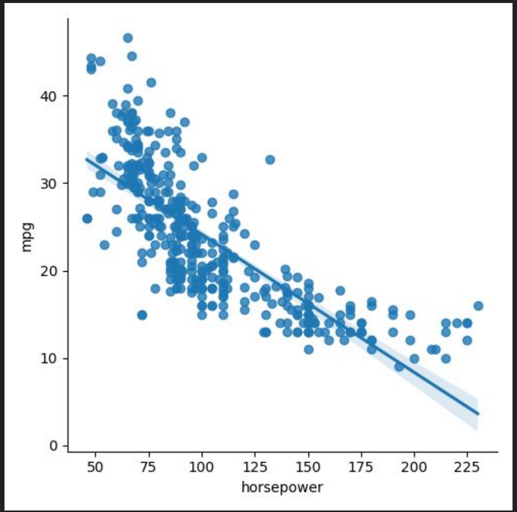
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-70.0117	6.645	-10.536	0.000	-83.076	-56.947
model_year	1.2300	0.087	14.080	0.000	1.058	1.402

Omnibus:	21.407	Durbin-Watson:	0.775
Prob(Omnibus):	0.000	Jarque-Bera (JB):	15.843
Skew:	0.387	Prob(JB):	0.000363
Kurtosis:	2.391	Cond. No.	1.57e+03

Might other variables explain the difference?

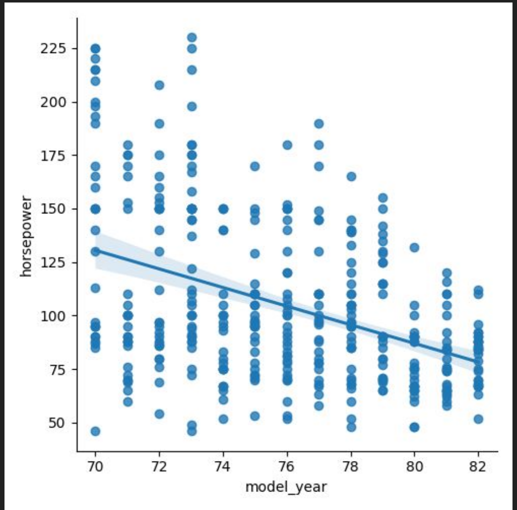
- Other variables are even stronger correlated with mpg. Horsepower for example is strongly negatively correlated with mpg. (R-squared = 0.606)





- We also note that model_year is somewhat negatively correlated with horsepower (R-squared = 0.173)

- Might our observed correlation between model_year and mpg be explained by the decrease in horsepower over time?



Does controlling for horsepower remove the correlation between model_year and mpg?

- No! When controlling for horsepower our coefficient drops from 1.23 to 0.66 and our t-value drops from 14.08 to 9.91.
- That is still a very high t-value with a rounded probability of 0.000 according to the table.
- We therefore conclude that the decrease in horsepower is not the only factor driving the increase in mpg over time.

```
results = smf.ols("mpg ~ model_year + horsepower", data=cars).fit()
results.summary()

```

✓ 0.1s

OLS Regression Results

Dep. Variable:	mpg	R-squared:	0.685
Model:	OLS	Adj. R-squared:	0.684
Method:	Least Squares	F-statistic:	423.9
Date:	Tue, 31 Jan 2023	Prob (F-statistic):	1.94e-98
Time:	13:50:21	Log-Likelihood:	-1134.5
No. Observations:	392	AIC:	2275.
Df Residuals:	389	BIC:	2287.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-12.7392	5.349	-2.382	0.018	-23.256	-2.223
model_year	0.6573	0.066	9.919	0.000	0.527	0.788
horsepower	-0.1317	0.006	-20.761	0.000	-0.144	-0.119

Omnibus:	11.834	Durbin-Watson:	1.054
Prob(Omnibus):	0.003	Jarque-Bera (JB):	12.068
Skew:	0.400	Prob(JB):	0.00240
Kurtosis:	3.316	Cond. No.	3.20e+03

Final thoughts

The analysis is based heavily on linear regression and that may not be the best model for change in fuel consumption over time. Perhaps a logistic model might be a more reasonable fit. It is however outside the scope of this assignment to explore other models.