Hur stor påverkan har en bils egenskaper på hur många mil per gallon den kan köra

Introduktion

I denna rapport skall jag, med hjälp av Python, undersöka och analysera data settet ”auto-mpg”. Data settet inkluderar namn på bilmodellen samt dess cylindrar, motorvolym, hästkrafter, vikt, acceleration, årsmodell, tillverkningsland samt mpg (miles per gallon). Med hjälp av denna data ska jag undersöka hur de olika egenskaperna på en bil kan påverka hur många mpg den kan köra. Därmed kan vi även dra slutsatser till hur man kan tillverka mer klimatsmarta bilar.

Metod

Jag kommer att använda mig utav python-paketen numpy, pandas, matplotlib, scipy, statsmodels.

Pandas kommer att användas för data behandling. För att läsa in det stora data settet och anpassa datan för att visa korrekt och relativ data. Jag kommer att gruppera datan och skapa ny tabell med relativ data såsom medelvärde och median. Datan kommer att grupperas genom tillverkningsland och årsmodell, främst.

Matplotlib kommer att användas för att plotta graferna. Det kommer för det mesta bara vara punktdiagram, med relativa titlar och axelnamn för att enkelt visa hur grafen ser ut.

Scipy och statsmodels kommer vara mina främsta verktyg för att modellera och analysera statistisk data från den framgivna datan. Efter att ha anpassat datat genom pandas kommer jag att använda dessa pakets olika funktioner för att få fram, exempelvis, konfidensintervall och t-disribution bland annat.

Analys

Efter att ha initialiserat data settet skapade jag en två nya tabeller. Ena tabellen var grupperad genom tillverkningsland och andra tabellen var grupperad genom årsmodell. Dessa två tabeller innehåller medianen, medelvärdet och standardavvikelse för mpg, hästkrafter, vikt och acceleration. Detta var gjort med kodsnutt 1.

Resultatet blev dessa två tabeller:

Tabell 1. Medelvärde, median och standardavvikelse (std) för mpg, motorvolym, hästkrafter, vikt och acceleration per tillverkningsland.

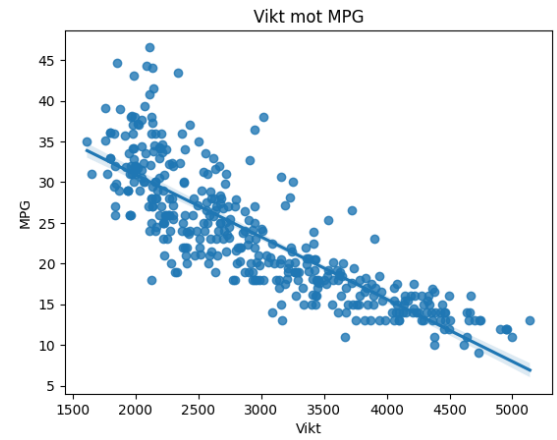
	MPG			**Motorvolym**			**Hästkrafter**			**Vikt**			**Acceleration**		
Land	Med-elvär-de	Med-ian	std	Med-elvär-de	Med-ian	std	Med-elvär-de	Med-ian	std	Med-elvär-de	Med-ian	std	Med-elvär-de	Med-ian	std
Europa	27.60	26.00	6.580	109.6	105.0	22.69	80.56	76.50	20.16	2433	2240	491.8	16.79	15.60	3.088
Japan	30.45	31.60	6.090	102.7	97.00	23.14	79.84	75.00	17.82	2221	2155	320.5	16.17	16.40	1.955
USA	20.03	18.50	6.440	247.5	250.0	98.38	119.0	105.0	39.90	3372	3381	795.3	14.99	15.00	2.736

Tabell 2. Medelvärde, median och standardavvikelse (std) för mpg, motorvolym, hästkrafter, vikt och acceleration per årsmodell.

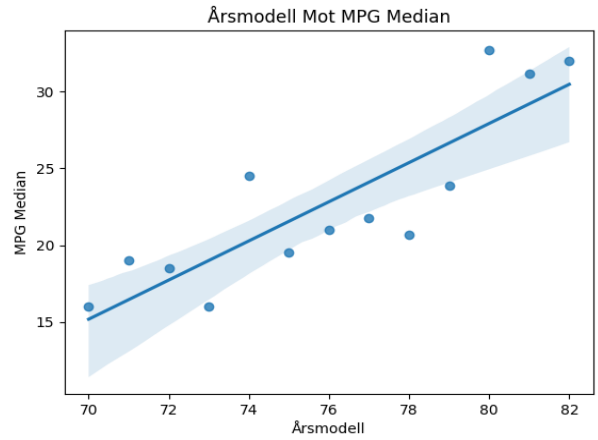
	MPG			**Motorvolym**			**Hästkrafter**			**Vikt**			**Acceleration**		
År	Med-elvär-de	Med-ian	std	Med-elvär-de	Med-ian	std	Med-elvär-de	Med-ian	std	Med-elvär-de	Med-ian	std	Med-elvär-de	Med-ian	std
70	17.69	16.00	5.339	281.4	307.0	124.4	147.8	150.0	53.73	3373	3449	852.9	12.95	12.50	3.331
71	21.11	19.00	6.676	213.9	232.0	115.2	107.0	95.00	38.57	3031	2962	1065	15.00	14.50	2.605
72	18.71	18.50	5.436	218.4	131.0	123.8	120.2	104.5	41.12	3238	2956	974.5	15.13	14.50	2.850
73	17.10	16.00	4.700	256.9	276.0	121.7	130.5	129.5	46.41	3419	3339	974.8	14.31	14.00	2.754

74	22.77	24.50	6.538	170.7	121.0	94.26	94.23	88.00	29.69	2878	2470	968.1	16.17	16.00	1.714
75	20.27	19.50	4.941	205.5	228.0	87.67	101.1	97.00	26.58	3177	3099	765.2	16.05	16.00	2.472
76	21.57	21.00	5.889	197.8	184.0	94.42	101.1	93.50	32.43	3079	3172	821.4	15.94	15.50	2.801
77	23.38	21.75	6.676	191.4	143.0	107.8	105.1	97.50	36.10	2997	2748	912.8	15.44	15.65	2.273
78	24.06	20.70	6.898	177.9	159.5	76.01	99.69	97.00	28.44	2862	2910	626.0	15.81	15.75	2.130
79	25.09	23.90	6.794	206.7	183.0	96.31	101.2	90.00	28.46	3055	3190	747.9	15.81	15.00	2.953
80	33.80	32.70	6.886	116.1	107.0	34.16	77.48	75.00	18.17	2442	2335	422.9	17.02	16.50	2.885
81	30.19	31.15	5.635	136.6	119.5	59.06	81.04	75.50	18.11	2530	2438	541.9	16.33	16.30	2.231

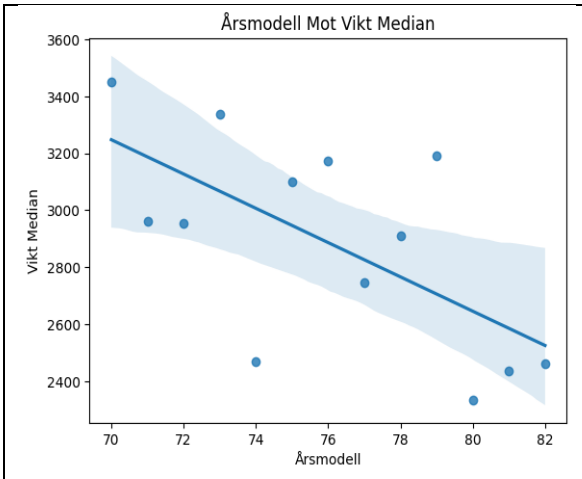
Därefter plottade jag fyra grafer. På den första grafen ser vi en negativ korrelation för vikt mot mpg. Ju högre vikten blir, desto lägre blir mpg. Den andra grafen visar inte mycket data, men med den datan vi ser så syns det att medianen för mpg har ökat med åren. Därefter på den tredje grafen ser vi att vikten har sjunkit med åren. Detta kan ha haft en påverkan på mpg som sett på tidigare graf. Tillsist, den fjärde grafen visar en motorvolym som sänkts med åren. Detta kan även haft en påverkan på mpg såsom synts på graf 2.



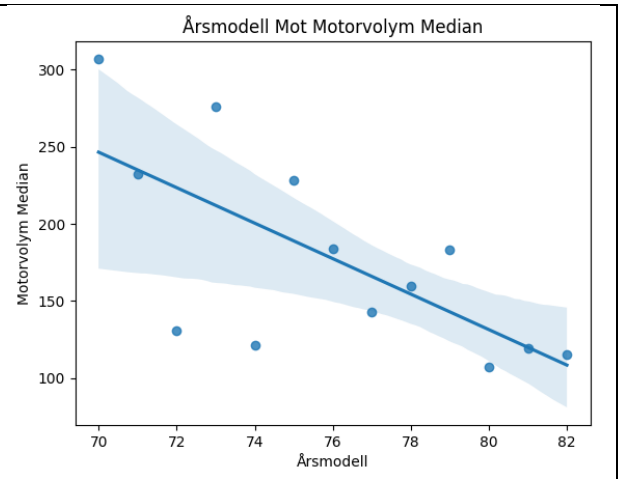
Graf 1. Vikten plottat mot mpg för alla bilar.



Graf 2. Median mpg för alla bilar för just det året plottat mot året.



Graf 3. Median vikten för alla bilar för just det året plottat mot året.



Graf 4. Median motorvolymen för alla bilar för just det året plottat mot året.

Efter det började jag göra ett hypotestest. Testet föreslår att med 95% konfidensgrad kunna säga att en bil skapad i USA, Europa, eller Japan ska ha en mpg på över 25. Därmed såg nollhypotesen ut på detta vis: H_0: $\mu \leq 25$, H_A: $\mu > 25$. Jag använde scipys ”scipy.stats.ttest_1samp()” funktion för att räkna ut ett p-värde. Detta visas i kodsnutt 3.

Tabell 3. T-test resultat från scipy.stats.ttest_1samp() funktionen för var tillverkningsland.

Land	**Resultat**
Europa	TtestResult(statistic=3.2619769955784466, pvalue=0.000871113768444249, df=67)
USA	TtestResult(statistic=-12.070475803378931, pvalue=1.0, df=244)
Japan	TtestResult(statistic=7.954992219947488, pvalue=5.750845201179692e-12, df=78)

Utifrån dessa resultat kan vi se de olika p-värden som vi kan använda för att jämföra med våran signifikans. Vi hade en signifikans på 0.05. Både europa och japan hade ett p-värde som var lägre än signifikansen, men USA hade inte det. Därmed kan vi förkasta nollhypotesen för europa och japan och säga att med 95% konfidens grad har de ett medelvärde för mpg högre än 25. Hur som helst kan vi inte förkasta nollhypotesen för USA, och kan därmed ej dra någon slutsats utav det.

Därefter började jag räkna ut konfidensintervall för mpg för var tillverkningsland. Detta gjorde jag med hjälp av kodsnutt 2. Jag initialiserade en ny dataframe där jag tog ut alla bilar med det specificierade tillverkningslandet. Därefter räknade jag ut stickprovsmedelvärdet och medelvärdesstandardfel med hjälp av numpy. Signifikansen (α) satte jag till 0.05 och därefter räknade jag ut konfidensintervallet med hjälp av scipys ”scipy.stats.t.interval()” funktion.

Tabell 4. Konfidensintervall för var tillverkningsland.

Land	**Resultat**
Europa	Confidence interval (α = 0.05) for average mpg of European cars: 26.0-29.2
USA	Confidence interval (α = 0.05) for average mpg of American cars: 19.2-20.8
Japan	Confidence interval (α = 0.05) for average mpg of Japanese cars: 29.1-31.8

Med dessa resultat kan vi se att Japanska bilar är de bästa när det gäller mpg, med Europeiska bilar strax efter. Däremot är amerikanska bilar sämst och ligger långt efter.

Slutsats

Med hjälp av analysen kan vi dra slutsatserna att amerikanska bilar har högre hästkrafter, motorvolym och vikt, men lägre mpg. Däremot har japanska och europeiska bilar lägre hästkrafter, motorvolym och väger mindre, och har därmed en högre mpg. Vi ser en tydlig relation mellan vikt mot mpg och motorvolym mot mpg och därmed kan vi dra slutsatsen att det är möjligt att dessa faktorer har en påverkan på mpg. Ifall vi vill få ut en högre mpg skulle det vara klokt att fokusera på att förbättra dessa faktorer.

Kodsnuttar

1.

```
mean_and_median_per_country = cars.groupby('origin').agg({'mpg': ['mean',
'median'], 'horsepower': ['mean', 'median'], 'weight': ['mean', 'median'],
'acceleration': ['mean', 'median']})

median_per_year = cars.groupby('model_year').agg({'mpg': ['mean', 'median'],
'horsepower': ['mean', 'median'], 'weight': ['mean', 'median'], 'accelera-
tion': ['mean', 'median']})
```

2.

```
# Subsetta data för tillverkningsland
europe = cars[cars['origin'] == 'europe']
usa = cars[cars['origin'] == 'usa']
japan = cars[cars['origin'] == 'japan']

# Beräkna n för stickprovet
n_europe = len(europe['mpg'])
n_usa = len(usa['mpg'])
n_japan = len(japan['mpg'])

# Beräkna medelvärde för stickprovet
mean_europe = np.mean(europe['mpg'])
mean_usa = np.mean(usa['mpg'])
mean_japan = np.mean(japan['mpg'])

# Signifikansnivå
alpha = 0.05

# Beräkna medelvärdesstandardfel med SciPy
sem_europe = scs.sem(europe['mpg'])
sem_usa = scs.sem(usa['mpg'])
sem_japan = scs.sem(japan['mpg'])

# Beräkna konfidensintervall med SciPy
lower_europe, upper_europe = scs.t.interval(confidence=1-alpha, df=n_europe-1,
loc=mean_europe, scale=sem_europe)
lower_usa, upper_usa = scs.t.interval(confidence=1-alpha, df=n_usa-1,
loc=mean_usa, scale=sem_usa)
lower_japan, upper_japan = scs.t.interval(confidence=1-alpha, df=n_japan-1,
loc=mean_japan, scale=sem_japan)

# Printa konfidensintervall
print(f'Confidence interval (\u03B1 = 0.05) for average mpg of european cars:
{round(lower_europe, 1)}-{round(upper_europe, 1)}')
print(f'Confidence interval (\u03B1 = 0.05) for average mpg of american cars:
{round(lower_usa, 1)}-{round(upper_usa, 1)}')
print(f'Confidence interval (\u03B1 = 0.05) for average mpg of japanese cars:
{round(lower_japan, 1)}-{round(upper_japan, 1)}')
```

3.

```
# Välj ut kolonnen 'mpg' som 'sample'.
sample_europe = europe['mpg']
sample_usa = usa['mpg']
sample_japan = japan['mpg']

# Beräkna stickprovsmedelvärde
xbar_europe = sample_europe.mean()
xbar_usa = sample_usa.mean()
xbar_japan = sample_japan.mean()

# Sätt mu till värde för nollhypotesen
mu = 25

# Genomför ensidigt t-test.
result_europe = scs.ttest_1samp(a=sample_europe, popmean=mu, alterna-
tive='greater')
result_usa = scs.ttest_1samp(a=sample_usa, popmean=mu, alternative='greater')
result_japan = scs.ttest_1samp(a=sample_japan, popmean=mu, alternative='great-
er')

# Printa resultatet
print(result_europe)
print(result_usa)
print(result_japan)
```