Exam 2

In [1]:
```
#Instructions
#Note that each question is for 1 point
#Exam total of 40 points possible

#Add your name and ID below
#Remember to save your work
#Run the setup sections and sanswer the questions that follow
```

In [2]:
```
# Student Name: Xingyu Chen

# Student ID: 110291925
```

In [3]:
```
#Loading Matplotlib and Seaborn

from matplotlib import pyplot as plt
import pandas as pd
import seaborn as sns
import numpy as np

sns.set_style ('darkgrid')
```

In [4]:
```
#Run the following cell.
#Use the following data set for Questions 1 - 5 below

import pandas as pd
Cars93_data = pd.read_csv("https://vincentarelbundock.github.io/Rdatasets/csv/MASS/Cars93.csv", index_col=0).dr
Cars93_data.head()
```

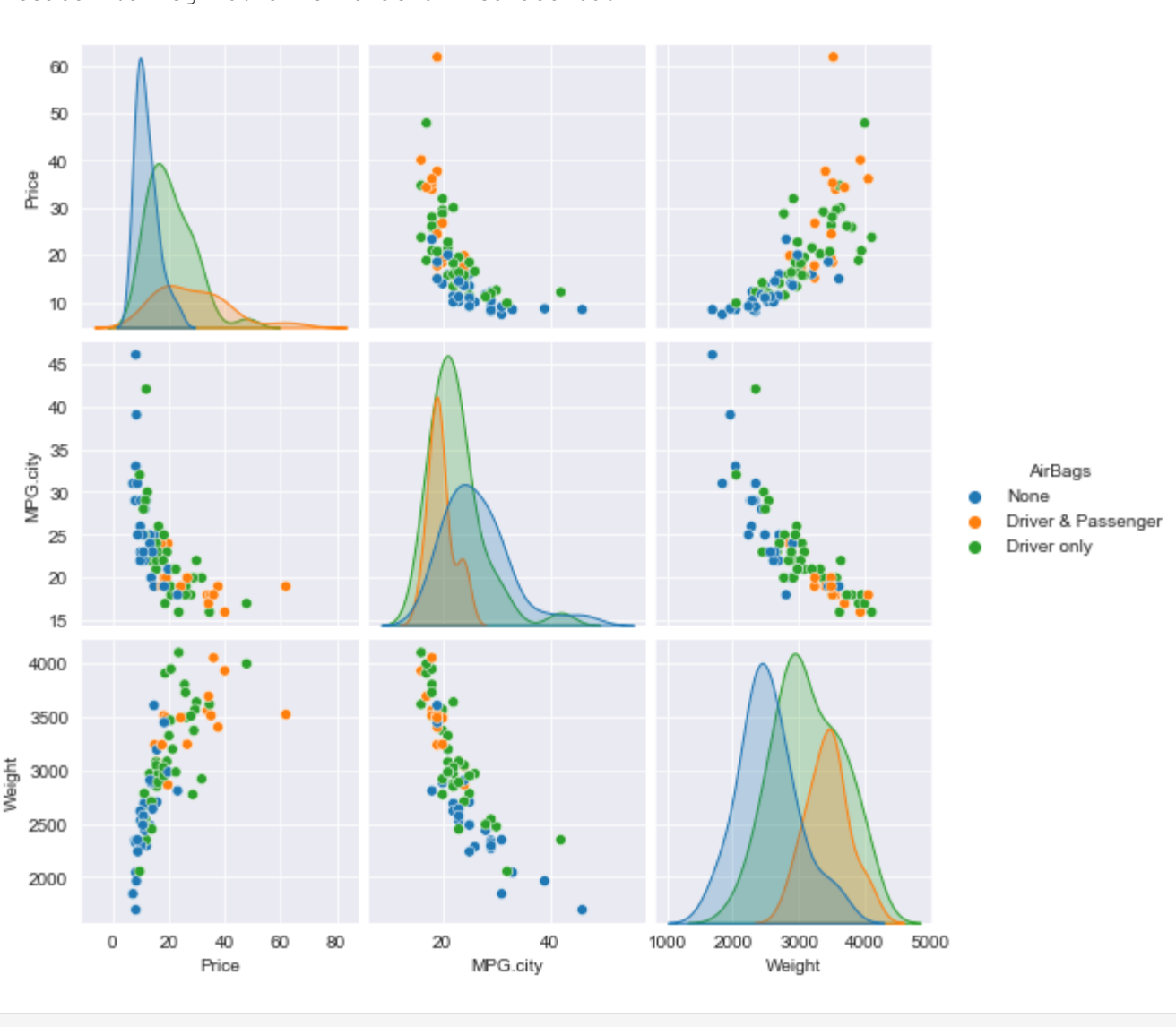
Out[4]:

	Manufacturer	Model	Type	Min.Price	Price	Max.Price	MPG.city	MPG.highway	AirBags	DriveTrain	...	Passengers	Lengt
1	Acura	Integra	Small	12.9	15.9	18.8	25	31	None	Front	...	5	17
2	Acura	Legend	Midsize	29.2	33.9	38.7	18	25	Driver & Passenger	Front	...	5	19
3	Audi	90	Compact	25.9	29.1	32.3	20	26	Driver only	Front	...	5	18
4	Audi	100	Midsize	30.8	37.7	44.6	19	26	Driver & Passenger	Front	...	6	19
5	BMW	535i	Midsize	23.7	30.0	36.2	22	30	Driver only	Rear	...	4	18

5 rows × 27 columns

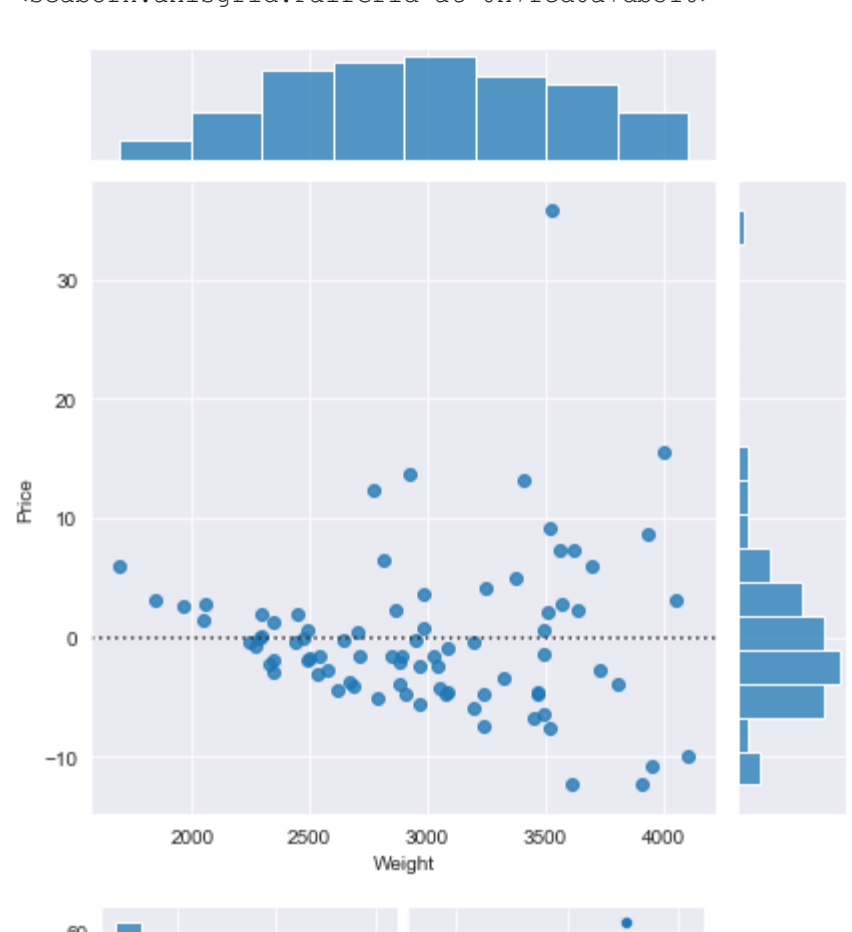
In [5]:
```
#Question 1.

#Run this code to display the pairplot for features Price, MPG.city and Weight

sns.pairplot(Cars93_data, hue='AirBags', vars=['Price','MPG.city','Weight'])

#Write Seaborn code in the next cell that shows that a non-linear model
# is probably not appropriate to model the Price-Weight relationship
```

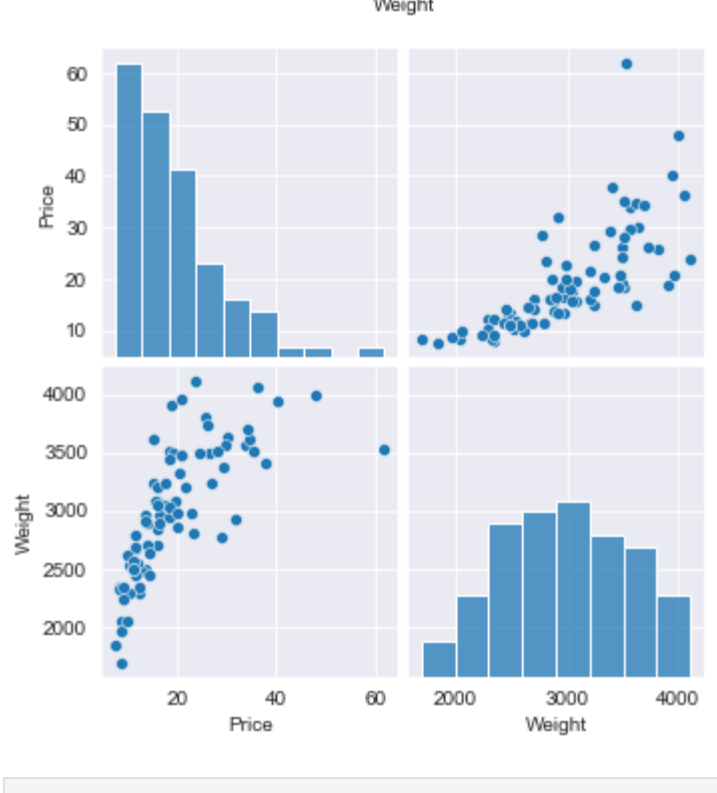
Out[5]: <seaborn.axisgrid.PairGrid at 0x7fcd193526d0>



In [95]:
```
#Write your answer for Question 1 below:
#Write Seaborn code in the next cell that shows that a non-linear model
# is probably not appropriate to model the Price-Weight relationship

sns.jointplot(x=Cars93_data.Weight, y=Cars93_data.Price,kind='resid')

sns.pairplot(Cars93_data, vars=['Price','Weight'])
```

Out[95]: <seaborn.axisgrid.PairGrid at 0x7fcd0a7db310>

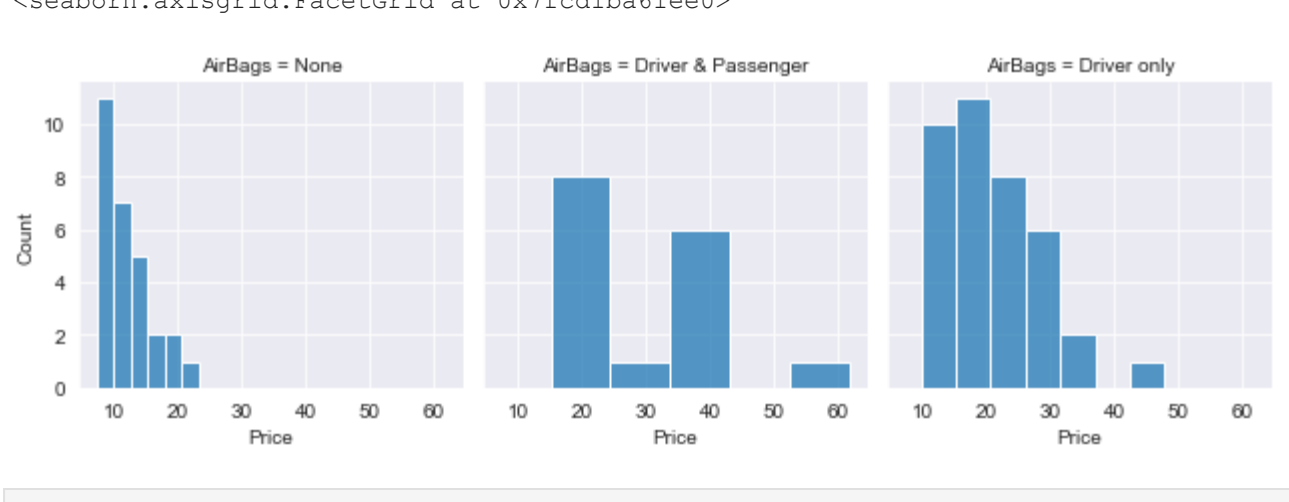




In [96]:
```
#Question 2:

#Explain your answer to Question 1 below:


# The pairplot suggests (to some extent) that a non-linear model would not be appropriate for the Price-Weight

# The residual plot for the same relationship appears to indicate that a linear model would be appropriate

# So a non-linear model is probably not appropriate to model the Price-Weight relationship
# a linear model is probably appropriate to model the Price-Weight relationship
```

In [15]:
```
#Question 3:

#Write Seaborn code to build a 1-row Facet Grid of boxplots for the Price feature using Cars93_data
#break out the chart by the AirBags feature

# Use .map_dataframe()
#Note you should have all lines of code in the same cell


#Write your answer below:

g=sns.FacetGrid(Cars93_data, col='AirBags')
g.map_dataframe(sns.histplot, x='Price')
```

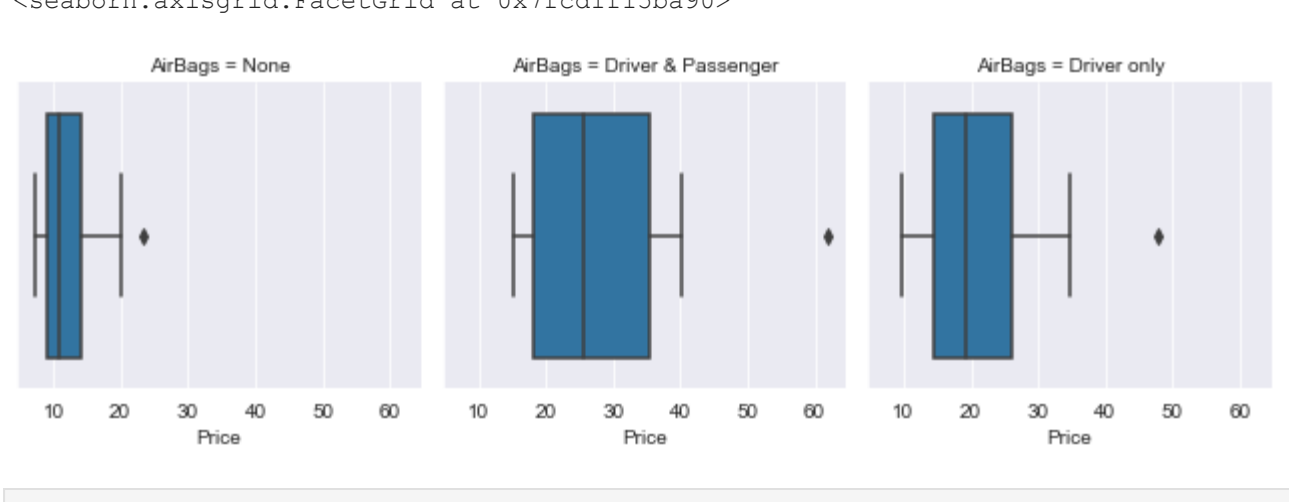
Out[15]: <seaborn.axisgrid.FacetGrid at 0x7fcd1ba61ee0>



In [16]:
```
#Question 4

#From your answer/chart in Question 3 above, which value of AirBags has
# the lowest 50th percentile for the feature Price?

#Write your answer below:
# AirBages = None
```

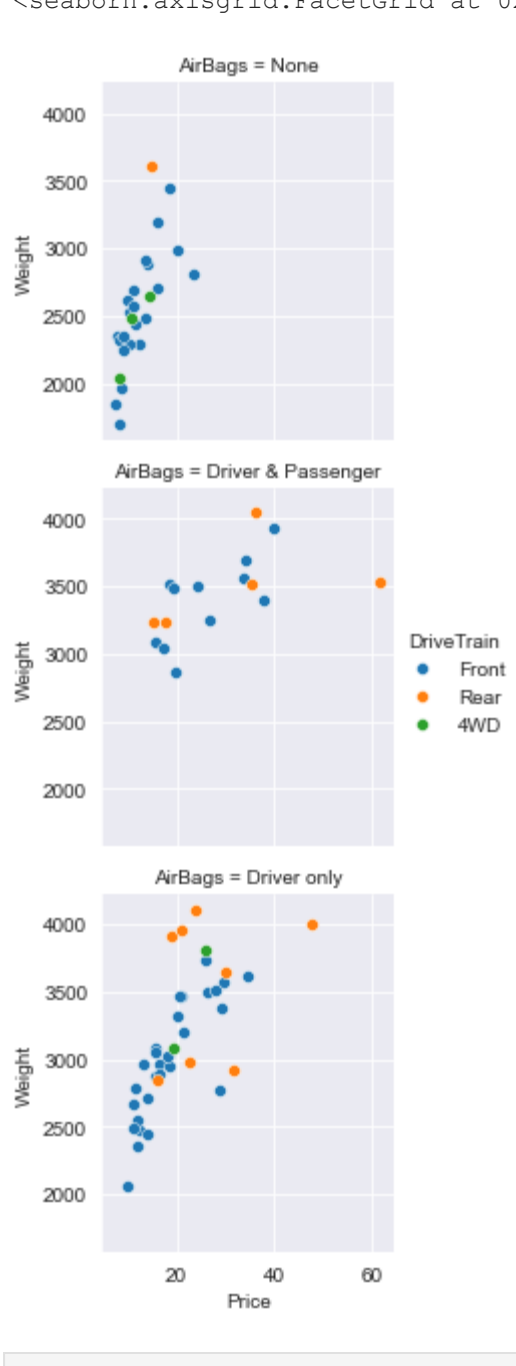
In [31]:
```
#Question 5

#From your answer to Question 3 above, which value of AirBags has the largest interquartile range (IQR)?

#Write your answer below:
g=sns.FacetGrid(Cars93_data, col='AirBags')
g.map_dataframe(sns.boxplot, x='Price')

# Airbages = Driver & Passenger
```

Out[31]: <seaborn.axisgrid.FacetGrid at 0x7fcd1f15ba90>



In [40]:
```
#Question 6

#Write Seaborn code to

#Write Seaborn code to build a 1-column Facet Grid of scatterplots using Cars93_data
#breakout the chart by the AirBags feature

#Add a scatterplot to the column for the variable Price
# Use .map_dataframe(). Put feature Price on the horizontal axis and Weight on the vertical axis

#Note you should have all lines of code in the same cell

#Use hue to breakout your chart with categorical variable DriveTrain

#Label the horizontal axis Price

#Label the vertical axis Weight

#Remember to pass hue to the sns.FacetGrid line and not the
#.map_dataframe() line

#Also add legend with g.add_legend

#Write your answer below:

g=sns.FacetGrid(Cars93_data, row='AirBags', hue = 'DriveTrain')
g.map_dataframe(sns.scatterplot, x='Price', y = 'Weight')
g.set_axis_labels('Price', 'Weight')
g.add_legend()
```

Out[40]: <seaborn.axisgrid.FacetGrid at 0x7fcd20284fa0>



In [41]:
```
#Question 7

# Using your chart from Question 6 above, when AirBags = None, the only dot for DriveTrain = Rear
# is between what range of values for variable Weight?

#Write your answer below:

# around 3550 weight, higgest value for variable weight, outlier
```

In [42]:
```
#Question 8

#Using your chart from Question 6 above, What can you infer about the Weight-Price relationship
# when feature DriveTrain = 4WD and feature AirBags = Driver & Passenger?

#Write your answer below:

# No green dot, there is no data when feature DriveTrain = 4WD and feature AirBags = Driver & Passenger
# so I can not infer about the Weight-Price relationship
```

In [43]:
```
#Question 9

#Using your chart from Question 6 above, what can you infer about the Weight-Price relationship
# when DriveTrain = Front and AirBags = Driver only?


#Write your answer below:

# They have positive relationship, when price increasing, the weight increasing
# meanwhile when price decreasing, the weight decreasing, when DriveTrain = Front and AirBags = Driver only
```

In [45]:
```
#Question 10

# From the chart in Question 1, what can you infer when AirBags = Driver only with respect to
# feature Weight and feature MPG.city?


#Write your answer below:

# when AirBags = Driver only, they are more dense for feature Weight and feature MPG.city.
```

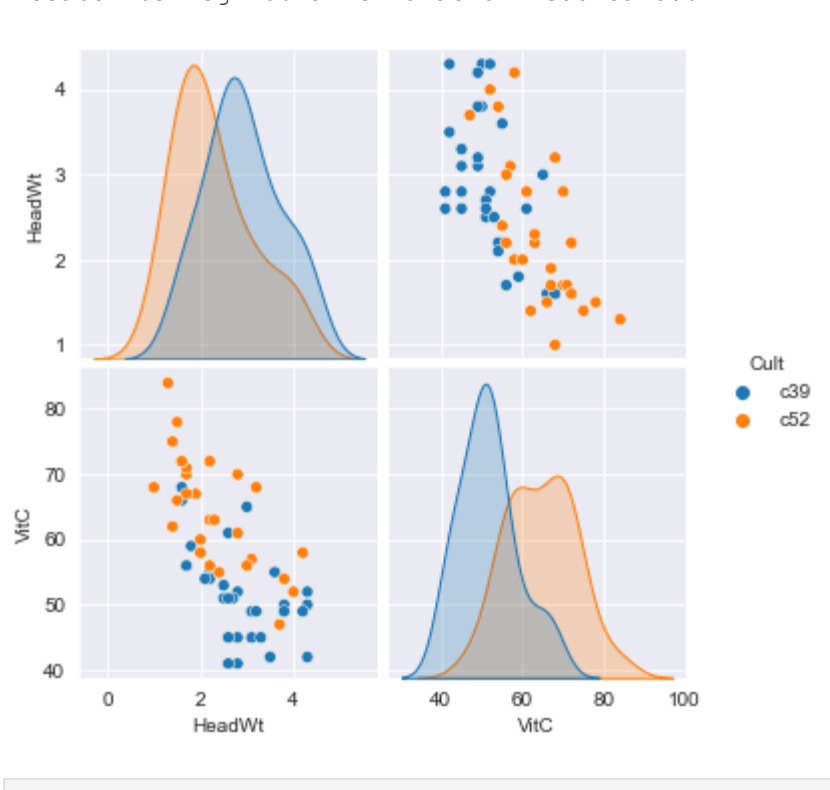
In [46]:
```
#Run the following cell.
#Use the following data set for Questions 11 - 20 below

import pandas as pd
import seaborn as sns
Bacteria_data = pd.read_csv("https://vincentarelbundock.github.io/Rdatasets/csv/MASS/cabbages.csv", index_col=0
Bacteria_data.head()
```

Out[46]:

	Cult	Date	HeadWt	VitC
1	c39	d16	2.5	51
2	c39	d16	2.2	55
3	c39	d16	3.1	45
4	c39	d16	4.3	42
5	c39	d16	2.5	53

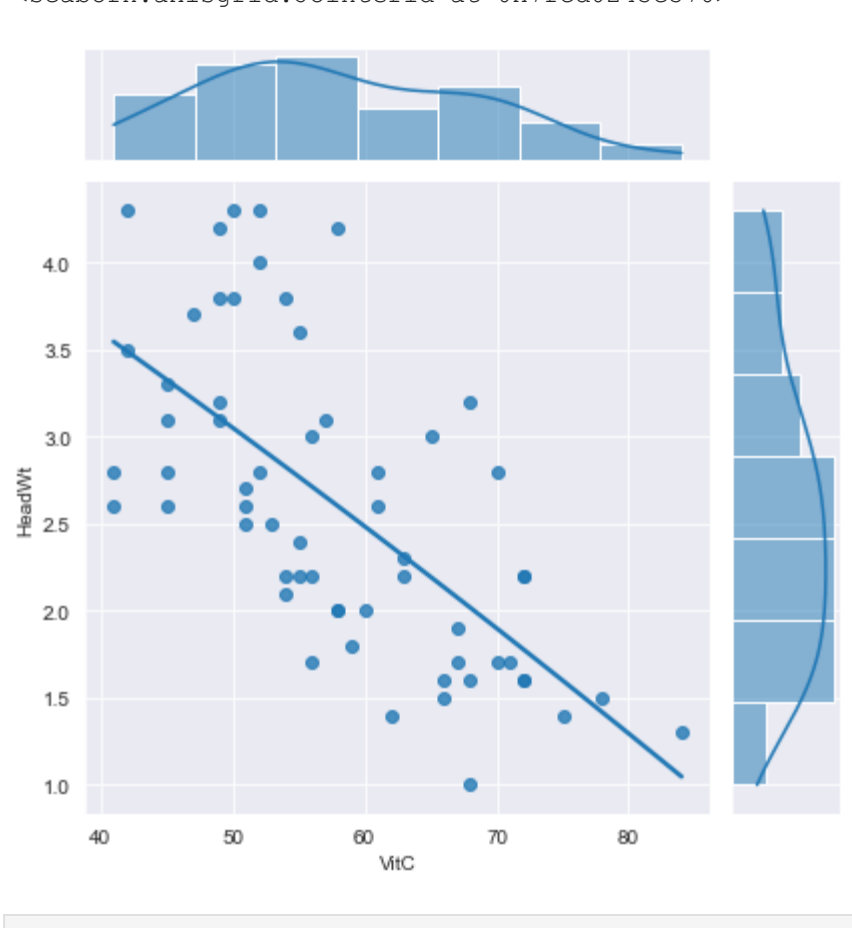
In [47]:
```
#Question 11

#Write the Seaborn code to render a pairplot on the entire dataset above
# Breakout the pairplot by the feature Cult

#Write your answer below

sns.pairplot(Bacteria_data, hue='Cult')
```

Out[47]: <seaborn.axisgrid.PairGrid at 0x7fcd01eaf580>



In [51]:
```
#Question 12

#From your chart in Question 12 above, write Seaborn code that would help you confirm
# whether or not a polynomial model would be appropriate to model the HeadWt-VitC relationship


#Write your answer below:

sns.jointplot(x=Bacteria_data.VitC, y=Bacteria_data.HeadWt,
            kind='reg',
            joint_kws={'ci': None, 'order': 2})
```

Out[51]: <seaborn.axisgrid.JointGrid at 0x7fcd0243e370>

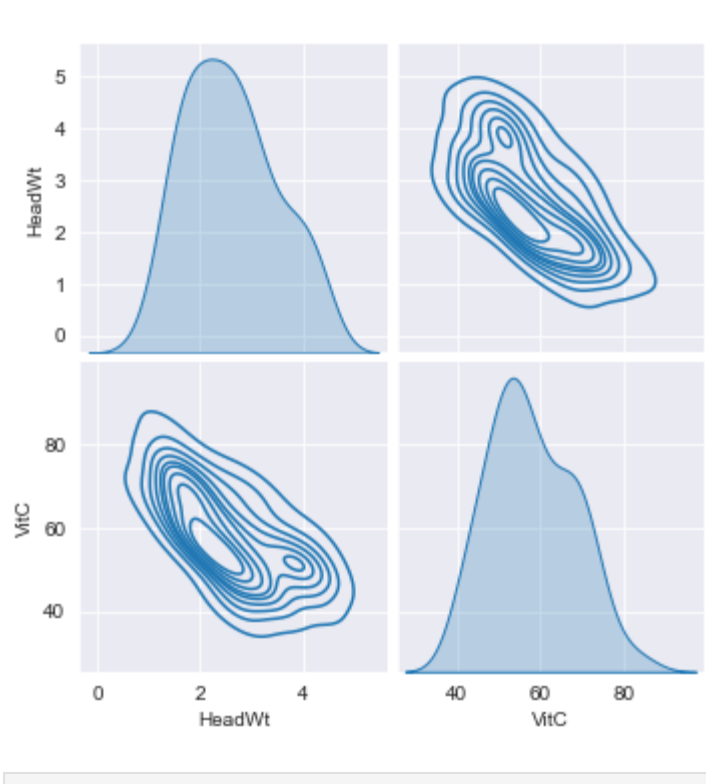


In [52]:
```
#Question 13

#From your answer in Question 12 above, what does your chart suggest about whether or not to use a polynomial m

#Write your answer below:



# a polynomial model would not be appropriate to model the HeadWt-VitC relationship
```

In [53]:
```
#Question 14

#What does your answer to Questions 11, 12 and 13 suggest about the HeadWt-VitC relationship?


#Write your answer below:

# they have negative linear relationship. when VitC increasing, the HeadWt decreasing
# meanwhile when VitC decreasing, the HeadWt increasing.
```

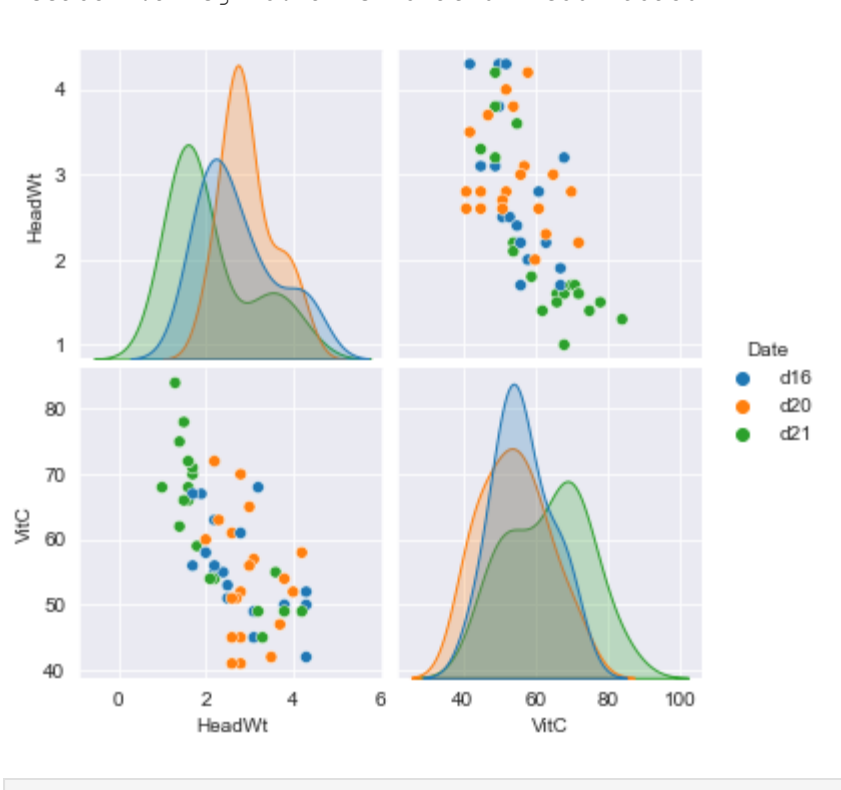
In [57]:
```
#Question 15

#Write Seaborn code to draw a pairplot (using kernel density plot) for the entire data frame Bacteria_data

#Write your answer below:
sns.pairplot(Bacteria_data, kind = 'kde')
```

Out[57]: <seaborn.axisgrid.PairGrid at 0x7fcd1f08cdc0>



In [59]:
```
#Question 16

#Write Seaborn code to draw a pairplot for the entire Bacteria_data data frame
#Breakout the plot by feature Date

#Write your answer below:
sns.pairplot(Bacteria_data, hue = 'Date')
```

Out[59]: <seaborn.axisgrid.PairGrid at 0x7fcd0429da30>

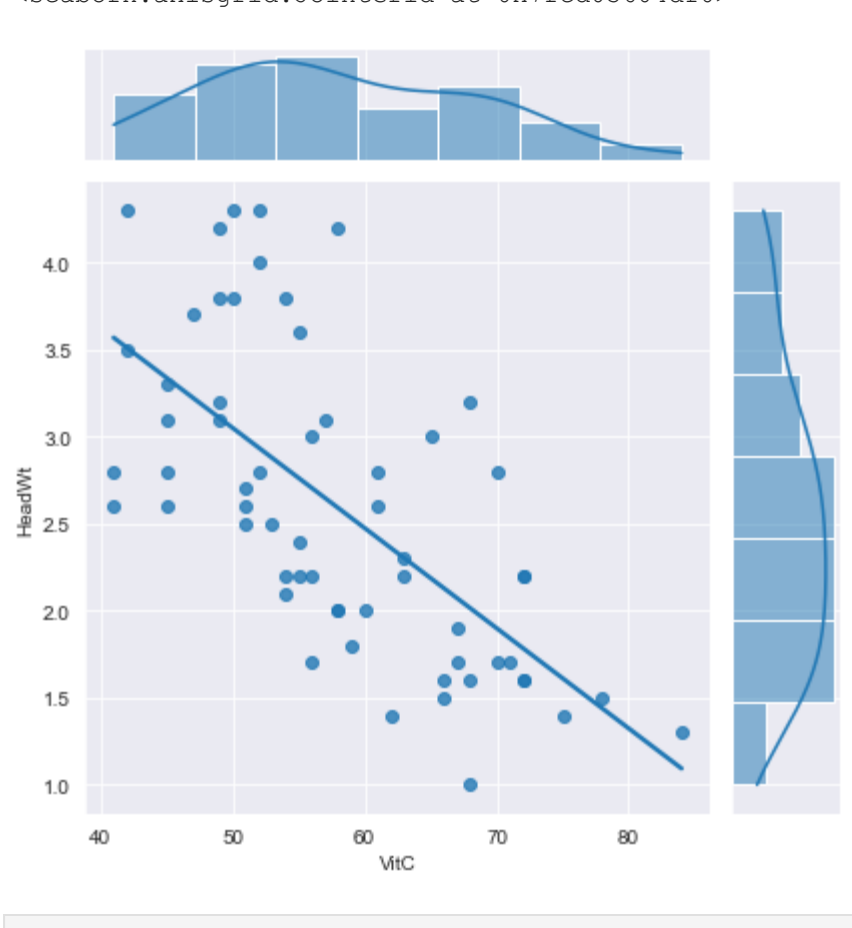


In [60]:
```
#Question 17

#From your chart in Question 16 above, what two values of feature Date can you conclude
# are mostly responsible for the deviation of the shape of the VitC feature
# from the benchmark normal distribution curve?

#Write your answer below:
# d21 and d16
```

In [ ]:
```
#Question 18

#From your chart in Question 16 above, what values of feature Date can you conclude
# are mostly responsible for the deviation of the shape of the HeadWt feature
# from the benchmark normal distribution curve?

#Write your answer below:

# d21 and d20
```

In [61]:
```
#Question 19

#Write Seaborn code to draw a regression plot for the HeadWt-VitC relationship
#Turn off the confidence interval

#Write your answer below:

sns.jointplot(x=Bacteria_data.VitC, y=Bacteria_data.HeadWt,
            kind='reg',
             joint_kws={'ci': None})
```

Out[61]: <seaborn.axisgrid.JointGrid at 0x7fcd03604df0>



In [62]:
```
#Question 20

#From your chart in Question 19 above, can an increase in VitC be expected
# to be associated with a decrease in HeadWt on average? (YES or NO)

#Write your answer below:

#YES
```

In [63]:
```
#Run this cell and use it to answer Questions 21 - 33 below

import pandas as pd
import seaborn as sns
import numpy as np
from matplotlib import pyplot as plt
import matplotlib
Insurance_data = pd.read_csv("https://vincentarelbundock.github.io/Rdatasets/csv/MASS/Insurance.csv", index_col
Insurance_data.head()
```

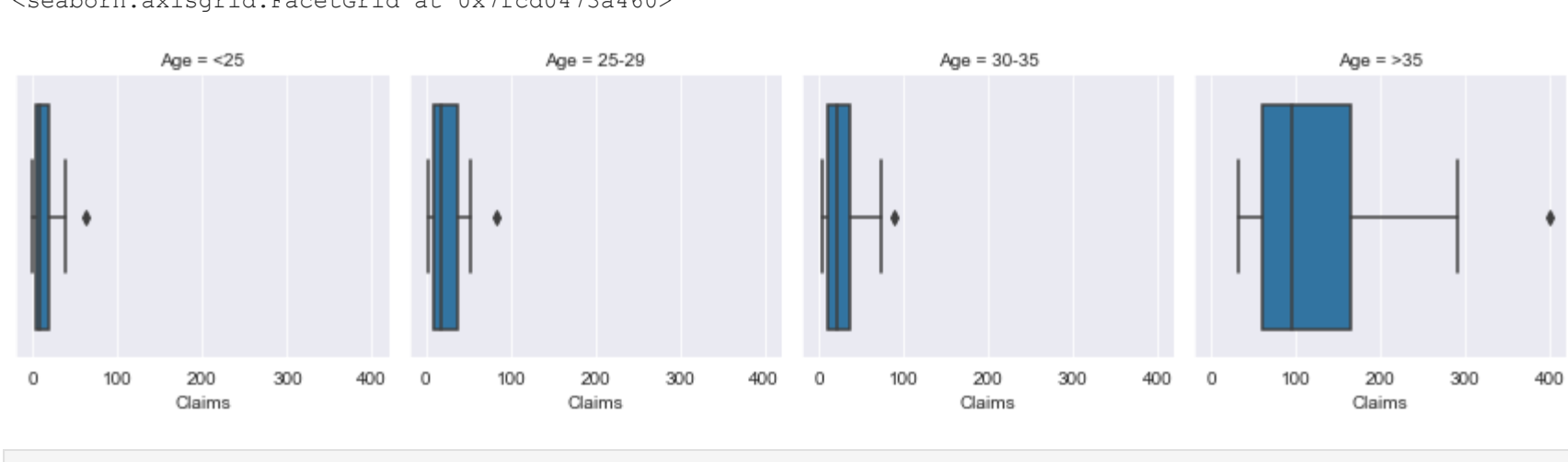
Out[63]:

	District	Group	Age	Holders	Claims
1	1	<1l	<25	197	38
2	1	<1l	25-29	264	35
3	1	<1l	30-35	246	20
4	1	<1l	>35	1680	156
5	1	1-1.5l	<25	284	63

In [64]:
```
##################################Variable definitions######################################

#District: factor: district of residence of policyholder (1 to 4): 4 is major cities.

#Group:an ordered factor: group of car with levels <1 litre, 1-1.5 litre, 1.5-2 litre, >2 litre.

#Age:an ordered factor: the age of the insured in 4 groups labelled <25, 25-29, 30-35, >35.

#Holders:numbers of policyholders.

#Claims:numbers of claims
```

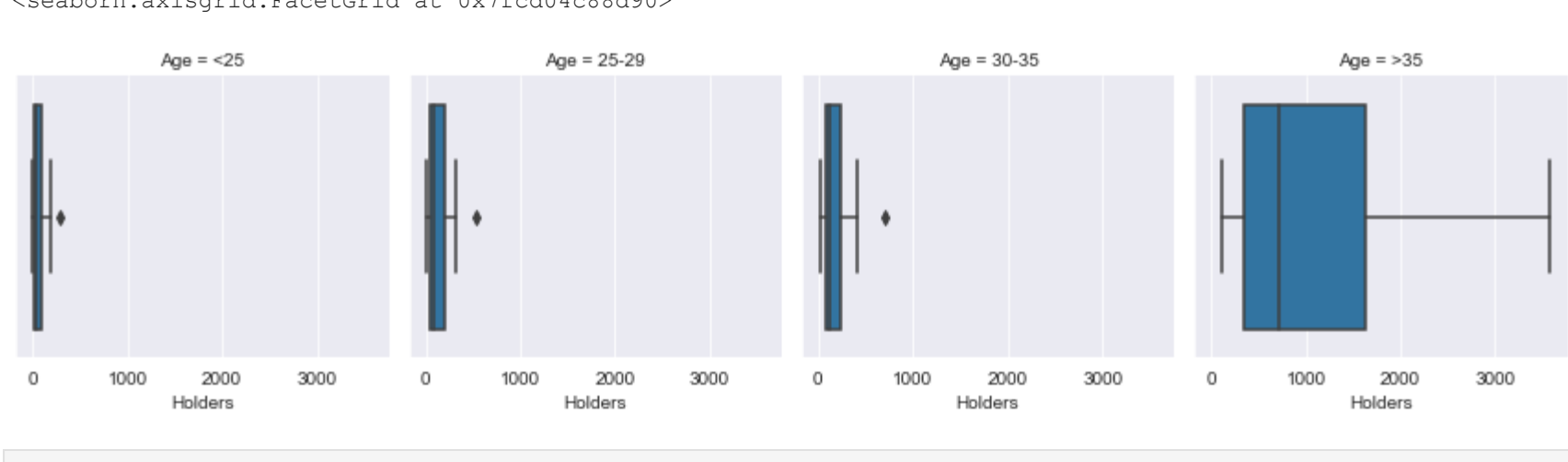
In [65]:
```
#Question 21

#Write Seaborn code to plot a FacetGrid of boxplots for the number of claims (Claims) feature, conditioned
# on the age of the insured (Age)?
#Put the charts on one row


#Write your code below:


g=sns.FacetGrid(Insurance_data, col='Age')
g.map_dataframe(sns.boxplot, x='Claims')
```

Out[65]: <seaborn.axisgrid.FacetGrid at 0x7fcd0473a460>



In [66]:
```
#Question 22

#Write Seaborn code to plot a FacetGrid of boxplots for the number of policyholders (Holders) feature, conditio
# on the age of the insured (Age)?
#Put the charts on one row


#Write your code below:

g=sns.FacetGrid(Insurance_data, col='Age')
g.map_dataframe(sns.boxplot, x='Holders')
```

Out[66]: <seaborn.axisgrid.FacetGrid at 0x7fcd04c88d90>



In [68]:
```
#Question 22

#What can you infer about the variability of observations for the Holders feature
# when Age is between 0 and 34 relative to when Age greater than 35?

#Write your answer below
# The range for the holder feature when age is between 0 and 34 is smaller than when Age greater than 35
# The IQR for the holder feature when age is between 0 and 34 is smaller than when Age greater than 35
# The Standard deviation for the holder feature when age is between 0 and 34 is smaller than when Age greater t
# The Variance for the holder feature when age is between 0 and 34 is smaller than when Age greater than 35
```

In [69]:
```
#Question 23

#What is responsible for your inference in Question 22 above?

#Write your answer below:
# The shape of boxplot
```

In [70]:
```
#Question 24

#What can you infer about the variability of observations of the Claims feature around their median
# as you read your chart in Question 21 from left to right?


#Write your answer below:
# The value of median is increasing from left to right
```

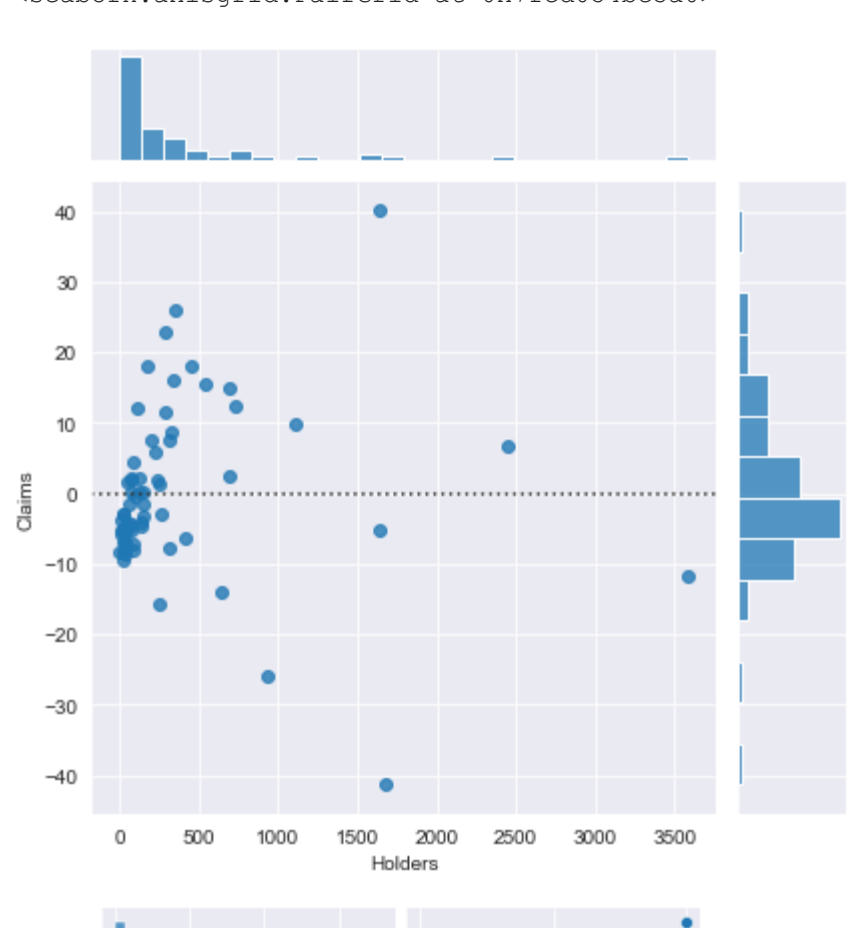
In [71]:
```
#Question 25

#Write Seaborn code to plot a FacetGrid of regression plots for the Claims-Holders relationship,
# conditioned on Age
#Put the charts on one row


#Write your code below:

g=sns.FacetGrid(Insurance_data, col='Age')
g.map_dataframe(sns.regplot, x='Holders', y='Claims')
```

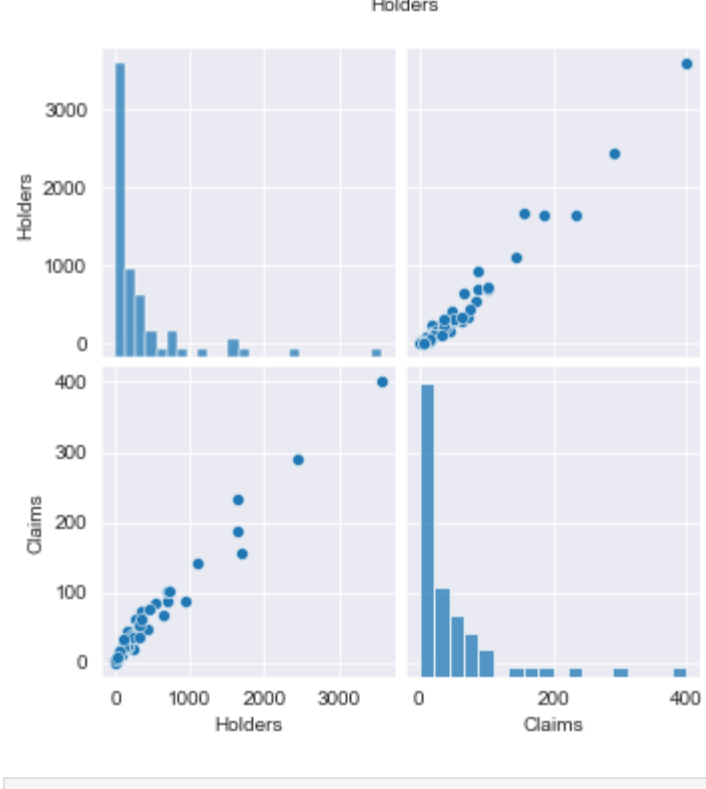
Out[71]: <seaborn.axisgrid.FacetGrid at 0x7fcd04eb7a90>

In [73]:
```
#Question 26

#Using your chart in Question 25 above, what can you infer about the Claims-Holders relationship overall?


#Write your answer below

# they have positive linear relationship, when holders increasing, the Claims increasing
# meanwhile when holders decreasing, the Claims decreasing
```

In [75]:
```
#Question 27

#Write Seaborn code to help you prove the type of model appropriate to use for the Claims-Holders relationship

#Write your answer below
# The pairplot suggests (to some extent) that a non-linear model would be appropriate for the Claims-Holders re

# The residual plot for the same relationship appears to indicate that a linear model would be appropriate


sns.jointplot(x=Insurance_data.Holders, y=Insurance_data.Claims,kind='resid')
sns.pairplot(Insurance_data, vars=['Holders','Claims'])
```

Out[75]: <seaborn.axisgrid.PairGrid at 0x7fcd054bc3a0>

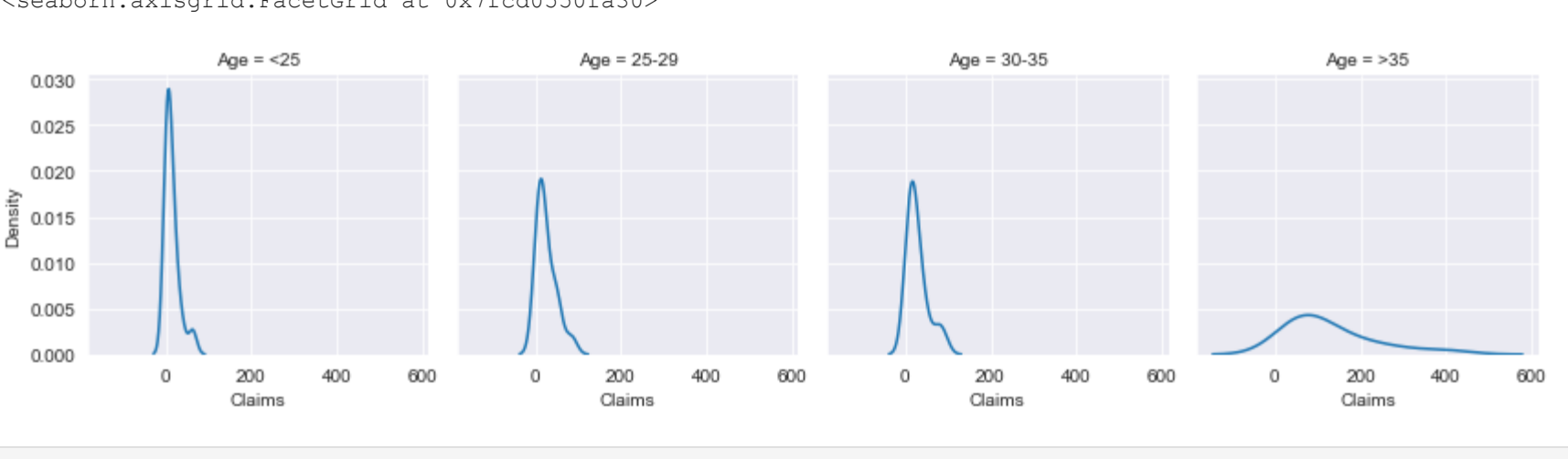




In [76]:
```
#Question 28

#From your answer to Question 27, is a non-linear model the appropriate model for the Claims-Holders relationsh


#Write your answer below
# NO
```

In [77]:
```
#Question 29

#Write Seaborn code to plot a FacetGrid of kde plots for the Claims feature,
# conditioned on Age
#Put the charts on one row


#Write your code below:
g=sns.FacetGrid(Insurance_data, col='Age')
g.map_dataframe(sns.kdeplot, x='Claims')
```

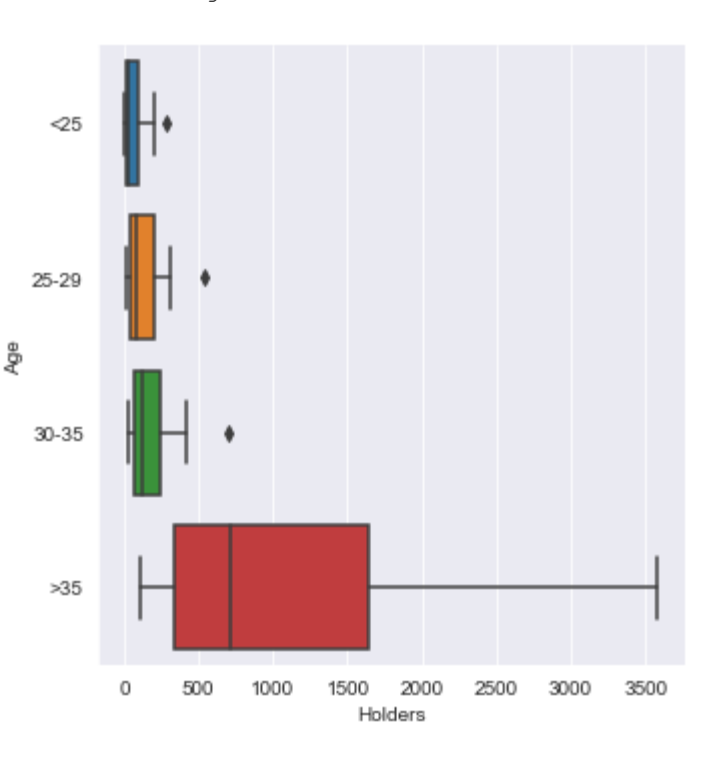
Out[77]: <seaborn.axisgrid.FacetGrid at 0x7fcd0550fa30>



In [78]:
```
#Question 30

#What does your chart in Question 29 suggest about the behaviour of Claims as you read from left to right
#(i.e as you consider changes in demographics indicated by the Age feature)?

#Write your answer below:

# probability of density of claims is decreasing fomr left to right meanwhile the median is increasing
```

In [80]:
```
#Question 31

#Write Seaborn code to generate a categorical plot of boxplots for the Holders feature
#conditioned on Age

#Write your answer below

sns.catplot(x='Holders', y = 'Age', data=Insurance_data, kind='box')
```

Out[80]: <seaborn.axisgrid.FacetGrid at 0x7fcd05d9aaf0>



In [81]:

```
#Question 32

#Assuming constant profitability of existing insurance policies across all Age groups.

#If you are the head of the insurance company sales team, and your current goal is to prevent as many
#policyholders as possible from cancelling their insurance policies, which age group would
# you most likely focus marketing resources on to keep their insurance policies active based on your chart in Q


#<25
#25-29
#30-35
#>35

#Write your answer below
# >35
```

In [82]:

```
#Question 33

#Justify your answer to Question 32 above.

#Write your answer below

# The age group greater than 35 has the most distribution so it is better focus on marketing on this age group
```

In [83]:

```
#Run this cell and use it to answer Questions 35 - 40 below

import pandas as pd
import seaborn as sns
import numpy as np
from matplotlib import pyplot as plt
import matplotlib
School_data = pd.read_csv("https://vincentarelbundock.github.io/Rdatasets/csv/mlmRev/bdf.csv", index_col=0).dro
School_data.head()

#Write your answer below
```

Out[83]:

	schoolNR	pupilNR	IQ.verb	IQ.perf	sex	Minority	repeatgr	aritPRET	classNR	aritPOST	...	currmeet	mixedgra	percmino	ar
1	1	17001	15.0	12.33333	0	N	0	14	180	24	...	1.83333	0	60	
2	1	17002	14.5	10.00000	0	Y	0	12	180	19	...	1.83333	0	60	
3	1	17003	9.5	11.00000	0	N	0	10	180	24	...	1.83333	0	60	
4	1	17004	11.0	10.00000	0	N	0	13	180	26	...	1.83333	0	60	
5	1	17005	8.0	6.66667	0	N	0	8	180	9	...	1.83333	0	60	

5 rows × 28 columns

In [44]:

```
# #################Variable Definitions###########################################################################

# schoolNR
# a factor denoting the school.

# pupilNR
# a factor denoting the pupil.

# IQ.verb
# a numeric vector of verbal IQ scores

# IQ.perf
# a numeric vector of IQ scores.

# sex
# Sex of the student.

# Minority
# a factor indicating if the student is a member of a minority group.

# repeatgr
# an ordered factor indicating if one or more grades have been repeated.

# aritPRET
# a numeric vector

# classNR
# a numeric vector

# aritPOST
# a numeric vector

# langPRET
# a numeric vector

# langPOST
# a numeric vector

# ses
# a numeric vector of socioeconomic status indicators.

# denomina
# a factor indicating of the school is a public school, a Protestant private school, a Catholic private school,

# schoolSES
# a numeric vector

# satiprin
# a numeric vector

# natitest
# a factor with levels 0 and 1

# meetings
# a numeric vector

# currmeet
# a numeric vector

# mixedgra
# a factor indicating if the class is a mixed-grade class.

# percmino
# a numeric vector

# aritdiff
# a numeric vector

# homework
# a numeric vector

# classsiz
# a numeric vector

# groupsiz
# a numeric vector
```

In [86]:

```
#Question 34

#Write Seaborn code to generate a FacetGrid of scatter plots for the IQ.perf-IQ.verb relationship according to

#Condition on the denomina feature first
#Condition on the Minority feature second
#Condition on the sex feature last

#Add legend

#Label the horizontal axis "Verbal IQ Score"
#Label the vertical axis "Overall IQ Score"

#You should have a 2-row X 4-column matrix of scatterplots

#Write your code below

g=sns.FacetGrid(School_data, col='denomina', row = 'Minority', hue='sex')
g.map_dataframe(sns.scatterplot, x='IQ.verb', y='IQ.perf')
g.set_axis_labels('Verbal IQ Score', 'Overall IQ Score')
g.add_legend()
```

Out[86]: <seaborn.axisgrid.FacetGrid at 0x7fcd079b3880>

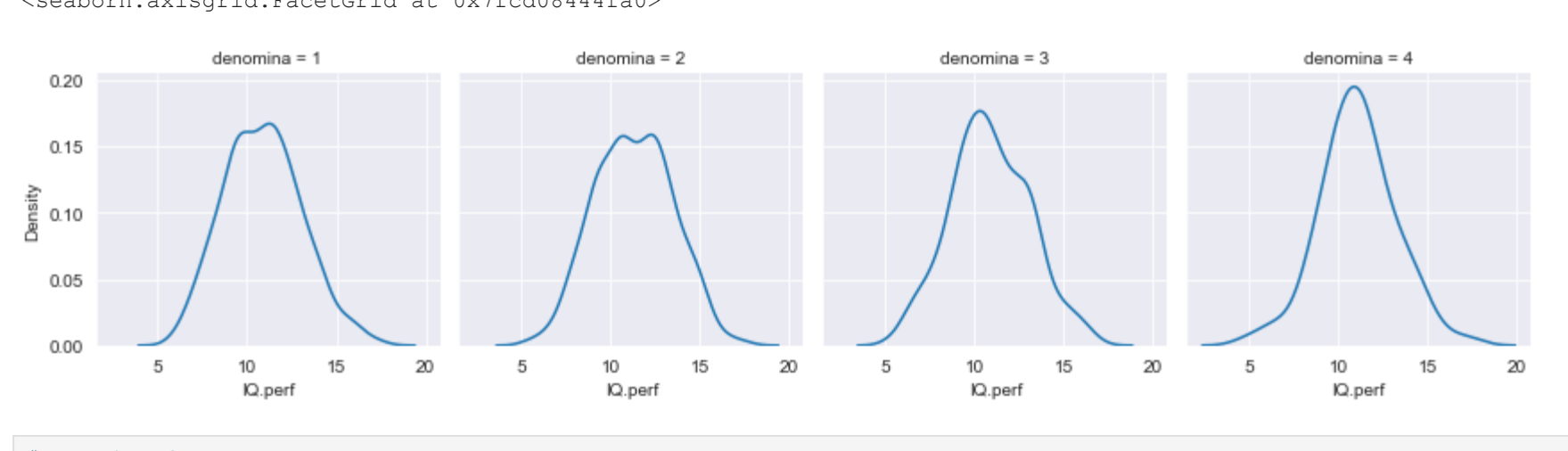
In [87]:

```
#Question 35

#Which of these charts represent a subset of data does your chart in Question 34 suggest you could safely delet
# from your dataset with minimal impact on your regression modeling results?

#Bottom rightmost scatterplot
#Top leftmost scatterplot
#Top rightmost scatterplot
#Bottom leftmost scatterplot


#Write your answer below
#Bottom rightmost scatterplot
```

In [88]:

```
#Question 36

#Write Seaborn code to generate a FacetGrid of kde plots for the IQ.perf feature according to the following:

#Condition on the denomina feature

#Add legend

#Write your code below



g=sns.FacetGrid(School_data, col='denomina')
g.map_dataframe(sns.kdeplot, x='IQ.perf')
g.add_legend()
```

Out[88]: <seaborn.axisgrid.FacetGrid at 0x7fcd08444fa0>



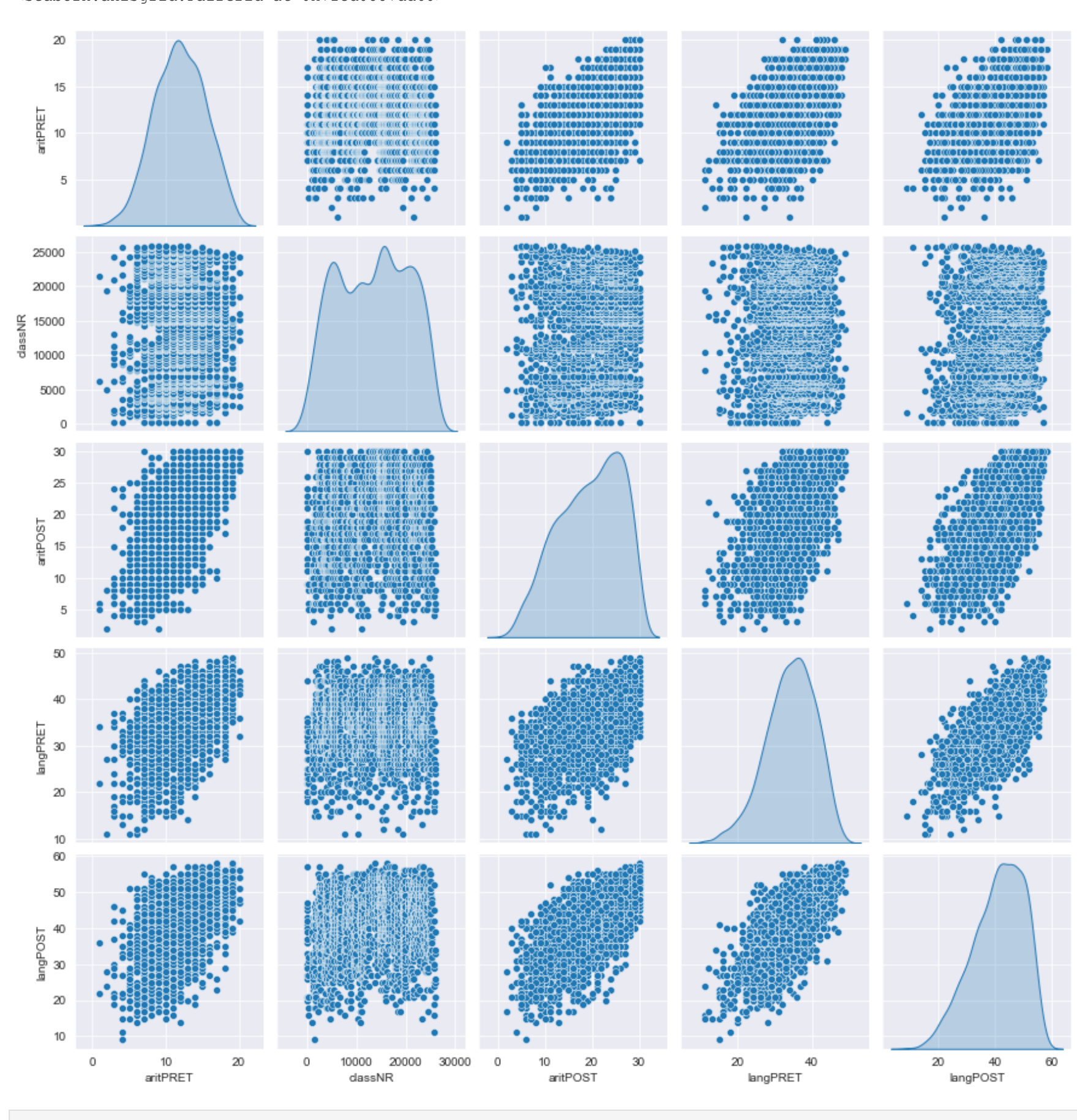
In [90]:

```
#Question 37

#For which value of the denomina feature is your chart in Question 36 closest to the benchmark
#standard normal deviation curve?

#Write your answer below

# denomina = 4
```

In [91]:

```
#Question 38

#Write Seaborn code to build a pairplot with kde plots in the diagonal using only the following variables:

#aritPRET

#classNR

#aritPOST

#langPRET

#langPOST

#Write your answer below

sns.pairplot(School_data, vars=['aritPRET','classNR','aritPOST','langPRET','langPOST'], diag_kind='kde')
```

Out[91]: <seaborn.axisgrid.PairGrid at 0x7fcd0887dd60>



In [92]:

```
#Question 39

#From your chart in Question 38, which of these relationships does not appear to exist?

#langPOST-aritPRET
#langPRET-classNR
#langPRET-langPOST
#aritPRET-langPRET


#Write your answer below

#langPRET-classNR
```

In [93]:

```
#Question 40


#Write Seaborn code to justify why a linear model would be appropriate for the langPOST-langPRET relationship

#Write your answer below


# The pairplot suggests (to some extent) that a non-linear model would be appropriate for the langPOST-langPRET

# The residual plot for the same relationship appears to indicate that a linear model would be appropriate


sns.jointplot(x=School_data.langPRET, y=School_data.langPOST,kind='resid')
```

Out[93]: <seaborn.axisgrid.JointGrid at 0x7fcd08335850>