**Module 2**

Here are a few helpful downloads for this module:

* [Video Transcripts](https://student.emeritus.org/courses/4765/files/2878716?wrap=1)
* [Download Video Transcripts](https://student.emeritus.org/courses/4765/files/2878716/download?download_frd=1)
* [Quick Reference Guide](https://student.emeritus.org/courses/4765/files/2878719?wrap=1)

Loc, scale

Dot . & Tab gives all available functions!

from scipy.stats import uniform

from scipy.stats import norm

import matplotlib.pyplot as plt

import numpy as np

Gaussian - normal - bell curve

mu\_X

**Coding\_Activity\_2.1**

dist1 = uniform(loc=10, scale=3)

dist1\_mean = dist1.mean()

dist1\_var = dist1.var()

dist1\_std = dist1.std()

x = np.linspace(9,14,100)

plt.plot(x,dist1.pdf(x))

p\_9 = dist1.pdf(9)

p\_11 = dist1.pdf(11)

p\_11\_or\_12\_ = dist1\_.pdf(11) + dist1\_.pdf(12)

p\_less\_than\_12 = dist1.cdf(12)

p\_between\_11\_and\_13 = dist1.cdf(13) - dist1.cdf(11)

p\_greater\_than\_12 = 1 - dist1.cdf(12)

dist1 = norm(loc=5, scale=10)

**for** i **in** range(1,501,1):

*#generate samples*

*#remember the random\_state*

samples = dist1.rvs(i,random\_state = 22)

*#find sample mean*

smean = np.mean(samples)

*#append mean to sample\_means*

sample\_means.append(smean)

ans3 = sample\_means[400] - 10 < .1

samples\_30\_or\_more = sample\_means[30:]

samples\_mean = np.mean(samples\_30\_or\_more)

samples\_std = np.std(samples\_30\_or\_more)

gauss\_dist = norm(loc = 5, scale = 10)

sample\_means\_gauss = []

**for** i **in** range(30,501,1):

*#generate samples*

*#remember the random\_state*

samples = gauss\_dist.rvs(i,random\_state = 22)

*#find sample mean*

smean = np.mean(samples)

*#append mean to sample\_means*

sample\_means\_gauss.append(smean)

gauss\_mean = np.mean(sample\_means\_gauss)

gauss\_standard\_deviation = np.std(sample\_means\_gauss)

nyc\_salary\_data = pd.read\_csv('data/nyc\_salaries.csv')

ans\_2 = nyc\_salary\_data['base\_salary'].mean()

ans\_3 = nyc\_salary\_data['base\_salary'].median()

first\_quartile = nyc\_salary\_data['base\_salary'].quantile(0.25)

third\_quartile = nyc\_salary\_data['base\_salary'].quantile(0.75)

iqr = third\_quartile - first\_quartile

lower = first\_quartile - 1.5\*iqr

upper = third\_quartile + 1.5\*iqr

salaries\_no\_outlier = nyc\_salary\_data.loc[((nyc\_salary\_data['base\_salary'] > lower) & (nyc\_salary\_data['base\_salary'] < upper))]

salaries\_no\_outlier = nyc\_salary\_data.loc[(nyc\_salary\_data['base\_salary']>(first\_quartile - 1.5\*iqr)) & (nyc\_salary\_data['base\_salary']<(third\_quartile + 1.5\*iqr))]

mean\_no\_outliers = salaries\_no\_outlier['base\_salary'].mean()

median\_no\_outliers = salaries\_no\_outlier['base\_salary'].median()

std\_numpy\_outliers = np.std(nyc\_salary\_data['base\_salary'])

std\_numpy\_no\_outliers = np.std(salaries\_no\_outlier['base\_salary'])

std\_pandas\_outliers = nyc\_salary\_data['base\_salary'].std()

std\_pandas\_no\_outliers = salaries\_no\_outlier['base\_salary'].std()

smolt\_mean = smolt['Reflectance'].mean()

smolt\_median = smolt['Reflectance'].median()

smolt\_std = smolt['Reflectance'].std()

smolt\_first\_quartile = smolt['Reflectance'].quantile(0.25)

smolt\_third\_quartile = smolt['Reflectance'].quantile(0.75)

first\_class = len(titanic.loc[titanic['class']=='First'])

first\_class\_over\_40 = len( titanic.loc[((titanic['class']=='First') & (titanic['age'] > 40))] )

p\_over\_40\_given\_first\_class = first\_class\_over\_40 / first\_class

second\_class = len(titanic.loc[titanic['class']=='Second'])

second\_class\_over\_40 = len( titanic.loc[((titanic['class']=='Second') & (titanic['age'] > 40))] )

p\_over\_40\_given\_second\_class = second\_class\_over\_40 / second\_class

first\_class = titanic.loc[titanic['class']=='First']['age']

second\_class = titanic.loc[titanic['class']=='Second']['age']

sns.histplot(first\_class)

sns.histplot(second\_class)

num\_survived = len( titanic.loc[(titanic['survived']==1)] )

survived\_over\_30 = len( titanic.loc[((titanic['survived']==1) & (titanic['age'] > 30))] )

p\_over\_30\_given\_survived = survived\_over\_30 / num\_survived

sns.histplot(titanic, x='age', hue='survived')

survived = **True**

**Quizes**

Uniform distributions model situations in which the outcomes are between two values, a and b, and all outcomes are equally probable. : True

When a fair, six-sided die is rolled, there are six possible outcomes: 1, 2, 3, 4, 5, and 6.

What is the probability of each outcome? : 1/6

What is the formula for the expected value E[X] for discrete and continuous uniform distributions? : (a+b)/2

The formula for the variance Var[X] for a discrete uniform distribution is (n2-1)/12. : True

What is the formula for the variance Var[X] of a continuous uniform distribution? : (b-a)2/12

To import uniform distribution packages from SciPy into Python, the statement used is

from scipy.stats import uniform -: True

*You are correct! The answer “True” is correct because to import uniform distribution packages from the library SciPy, we explicitly mention uniform type in the statement.*

A uniform distribution ranging from a to b is made using this Python statement: U = uniform(loc,scale).

What does the constructor loc represent? -: Loc = a

*You are correct! The answer “loc=a” is correct because the constructor loc in the function represents the starting range of the uniform distribution, so it is represented as “a”.*

A uniform distribution ranging from a to b is made using this Python statement:  U = uniform(loc,scale).

What does the constructor scale represent? -: scale=b-a

*You are correct! The answer “scale=b-a” is correct because the constructor scale in the function represents half of the width of the uniform distribution, so it is represented as “b-a”.*

If a uniform distribution is formulated in Python as U = uniform(loc=10, scale=5),

what is the output of the statement U.mean()? -: 12.5

*You are correct! The answer “12.5” is correct because the statement returns the mean of the uniform distribution. Here a=10, b=15, and the formula for the mean is (a+b)/2, which gives the output as “12.5”.*

If a uniform distribution is formulated in Python as U = uniform(loc=10, scale=5),

what does the statement U.var() represent? -: Variance

*You are correct! The answer “Variance” is correct because this function is used to get the variance of a uniform distribution.*

In Python, for a uniform distribution U the function U.rvs(size)

is used to sample from the distribution U. -: True

*You are correct! The answer “True” is correct because the function*rvs(size)*provided with the size constructor is used to provide samples from a distribution object of count “size”.*

The uniform distribution is often referred to as the “bell curve”. : False

*False because the graph of uniform distribution is a straight line, so it cannot be referred to as a bell curve. A normal distribution looks like a bell curve.*

The probability density function for normal distribution is

1

2

π

σ

2

e

−

(

x

−

μ

)

2

2

σ

2

True!

*True because the formula shown for normal distribution is correct.*

How do you represent the mean and variance of a random variable “**x̄n**”? -: Mean= μ**x̄n**

Variance= σ2**x̄n**

*You are correct! The answer “*Mean= μ**x̄n,**Variance= σ2**x̄n***” is correct because the symbol for mean and variance is “*μ” *and* “σ2”. *For the random variable* “**x̄n**” *the symbols are used together with the symbol of the variable for which the measures are used.*

The equation σ2**x̄n**=**(**σ2**x)**/n

states that the larger the sample size, the smaller the variance of the random variable **x̄n. -: True**

*You are correct! The answer “True” is correct because in the formula “*σ2**x̄n**=**(**σ2**x)**/n*”, the sample size “n” is in the denominator, which states that the increase in “n” will reduce the variance of the random variable* “**x̄n**”.

The Central Limit Theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population, then the distribution of the sample means will be approximately normally distributed. -: True

*You are correct! The answer “True” is correct because the Central Limit Theorem states that the sampling distribution of the mean approaches a normal distribution as the size of the sample increases.*

What is the general threshold decided by statisticians for random variable sample size? -: 30

*You are correct! The answer “30” is correct because this is the threshold decided by*statisticians for random variable sample size.

Which of the following is a multivariate random variable? : X=[D,L]

*You are correct! The answer “X=[D,L]” is correct because a multivariate random variable is a collection of random variables.*

The mean of a multivariate random variable is the scalar of the means of the individual components. : False

*You are correct! The answer “False” is correct because the mean of a multivariate random variable is the vector of the means of the individual components.*

What do you call the matrix that is used to compute the variances and covariances of a multivariate random variable? : Covariance matrix

*You are correct! The answer “Covariance matrix” is correct because it summarizes the variances and covariances of a set of vectors.*

What is the formula to calculate covariance? : Covar[X] = E[(X-E[X])(X-E[X])2]

*You are correct! The answer “*Covar[X] = E[(X-E[X])(X-E[X])2]*” is correct because X minus the expectation of X multiplied by X minus expectation of X whole squared is the formula for covariance.*

What does the covariance between variables “A” and “B” represent? :

The relationship between “B” and “A”

The relationship between “A” and “B”

*You are correct! The answers “*The relationship between “A” and “B”,” *and “*The relationship between “B” and “A”*” are correct because covariance is a measure of the relationship between two variables and the variance is the same, regardless of direction.*

Consider the following covariance matrix:

∑

=

[

|  |  |  |
| --- | --- | --- |
| **σ**  00  2 | **σ**  01  2 | **σ**  02  2 |
| **σ**  10  2 | **σ**  11  2 | **σ**  12  2 |
| **σ**  20  2 | **σ**  21  2 | **σ**  22  2 |

]

  The entries in the diagonal, “σ200”, “σ211” and “ σ222”, represent the variances of the variables 1, 2, and 3, respectively. : True

*You are correct! The answer “*True*” is correct because in a covariance matrix the values in the diagonal are the variances of the respective variables.*

The covariance of “X1” with “X2” is equal to the covariance of “X2” with “X1”. : True

*You are correct! The answer “*True*” is correct because the covariance* *matrix has an axis of symmetry along its diagonal, hence these are the same.*

In Python, import seaborn as sns is used for statistical graphics. : True

*You are correct! The answer “*True*” is correct because seaborn is a data visualization library built on top of matplotlib and therefore is used for statistical graphics.*

In Python, what would you use to generate a covariance matrix on a dataframe? : df.cov()

*You are correct! The answer “*df.cov()*” is correct because the function “*cov()*” is used to build a covariance matrix for a dataframe.*

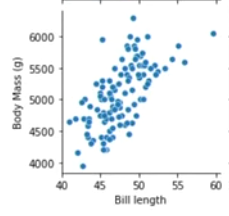
In Python what is the output of the given statement df.corr()? : Correlation matrix

*You are correct! The answer “*Correlation matrix*” is correct because the function “*corr()*” is used to build a correlation matrix for a dataframe in Python.*

In Python, to plot a correlation matrix, the function “sns()” is used. : False

*You are correct! The answer “*False*” is correct because to plot a correlation matrix the function “*sns.pairplot(df)*” is used. The “*sns*” itself is just the alias of the Python library seaborn.*

The plot below shows a weak correlation between the variable “Body Mass” and “Bill length”.



: False

*You are correct! The answer “*False*” is correct because the correlation between the variables*“Body Mass” *and*“Bill length” *is strong. The increase in the value of one variable results in an increase in the value of the second variable as well.*

What is the symbol used to represent correlation measures in a correlation matrix? : Rho,⍴

*You are correct! The answer “*rho,⍴*” is correct because the symbol is used for correlation measures.*

Consider the following correlation matrix:

C

o

r

r

[

X

]

=

[

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | **ρ**  1  ,  2 | … | **ρ**  1  ,  n |
| **ρ**  2  ,  1 | 1 | … | **ρ**  2  ,  n |
| ⋮ | ⋮ | ⋱ | ⋮ |
| **ρ**  n  ,  1 | **ρ**  n  ,  2 | … | 1 |

]

What is the correct formula to get the value of ⍴1,2? : ⍴1,2 = 𝜎1,22/𝜎1𝜎2

*You are correct! The answer “*⍴1,2 = 𝜎1,22/𝜎1𝜎2*” is correct because correlation “*⍴1,2*” is obtained by taking the 1,2 covariance and dividing it by the standard deviations of X1 and X2.*

A positive correlation means that when one goes up, the other tends to go down, and when one goes down, the other tends to go up. : False

*You are correct! The answer “*False*” is correct because a positive correlation means that they both go up and down together.*

What is the equation to get the conditional probability of variable Y given X? : Py(Y|X=x)= Px,y(X,Y)/Px(X)

*You are correct! The answer “*Py(Y|X=x)= Px,y(X,Y)/Px(X)*” is correct because* *the conditional probability of Y given X is equal to the joint pdf of X and Y divided by the marginal distribution of X.*

In Python statement

sns.histplot(data=gentoo, x=’Bill Length’, hue=’sex’),

the constructor “hue” is used to display bar heights. : False

*You are correct! The answer “*False*” is correct because “*hue*” is used as a semantic variable that is mapped to determine the color of plot elements.*

If two variables ‘X’ and ‘Y’ are independent of each other, then what is the joint distribution PX,Y(X,Y)? : PX,Y(X,Y) =PX(X)PY(Y)

*You are correct! The answer “*PX,Y(X,Y) =PX(X)PY(Y)*” is correct because if the variables are independent, their joint distribution is equal to the multiplication of their marginal distributions.*

If two variables ‘X’ and ‘Y’ are said to be independent, then

Py(Y|X=x1)=Py(Y|X=x2)=Py(Y). : True

*You are correct! The answer “*True*” is correct because the distribution of Y given X=x1 or the distribution of Y given X=x2 is equal to the marginal distribution of Y if the variables are independent.*

*Quiz*

If two variables ‘X’ and ‘Y’ are said to be independent, then

Py(Y|X=x1)=Py(Y|X=x2)=Py(Y). True

What does the covariance between variables “A” and “B” represent?

The relationship between “B” and “A”

The relationship between “A” and “B”

The uniform distribution is often referred to as the “bell curve”. False

The plot below shows a weak correlation between the variable “Body Mass” and “Bill length”.

False

In Python statement

sns.histplot(data=gentoo, x=’Bill Length’, hue=’sex’),

the constructor “hue” is used to display bar heights.

False

A uniform distribution ranging from a to b is made using this Python statement: U = uniform(loc,scale).

What does the constructor scale represent?

scale=b-a

In Python, what is the output of the given statement df.corr()?

Correlation matrix

What is the equation to get the conditional probability of variable Y given X?

Py(Y|X=x)= Px,y(X,Y)/Px(X)

For a correlation matrix

C

o

r

r

[

X

]

=

[

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | **ρ**  1  ,  2 | … | **ρ**  1  ,  n |
| **ρ**  2  ,  1 | 1 | … | **ρ**  2  ,  n |
| ⋮ | ⋮ | ⋱ | ⋮ |
| **ρ**  n  ,  1 | **ρ**  n  ,  2 | … | 1 |

]

The correct formula to get the value of ⍴1,2 is:

⍴1,2 = 𝜎1,22/𝜎1𝜎2

Uniform distributions model situations in which the outcomes are between two values, a and b, and all outcomes are equally probable.

True

The covariance of “X1” with “X2” is equal to the covariance of “X2” with “X1”.

True

Which of the following is a multivariate random variable?

X=[D,L]

A uniform distribution ranging from a to b is made using this Python statement U = uniform(loc,scale).

What does the constructor loc represent?

Loc = a

Consider the following covariance matrix:

∑

=

[

|  |  |  |
| --- | --- | --- |
| **σ**  00  2 | **σ**  01  2 | **σ**  02  2 |
| **σ**  10  2 | **σ**  11  2 | **σ**  12  2 |
| **σ**  20  2 | **σ**  21  2 | **σ**  22  2 |

]

  The entries in the diagonal, “σ200”, “σ211” and “ σ222”, represent the variances of the variables 1, 2, and 3, respectively.

True

The matrix that is used to compute the variances and covariances of a multivariate random variable is called

What is the formula to calculate covariance?

Covar[X] = E[(X-E[X])(X-E[X])2]

When a fair, six-sided die is rolled, there are six possible outcomes: 1, 2, 3, 4, 5, and 6.

What is the probability of each outcome?

1/6

How do we represent the mean and variance of a random variable “**x̄n**”?

Mean= μ**x̄n**

Variance= σ2**x̄n**

The Central Limit Theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population, then the distribution of the sample means will be approximately normally distributed.

True

What is the symbol used to represent correlation measures in a correlation matrix?

Rho,⍴

The formula for the variance Var[X] for a discrete uniform distribution is (n2-1)/12.

True

A positive correlation means that when one goes up, the other tends to go down, and when one goes down, the other tends to go up.

False

To import uniform distribution packages from SciPy into Python, the statement used is

from scipy.stats import uniform

True

The equation σ2**x̄n**=**(**σ2**x)**/n

states that the larger the sample size, the smaller the variance of the random variable **x̄n.**

**True**

What is the general threshold decided by statisticians for random variable sample size?

30

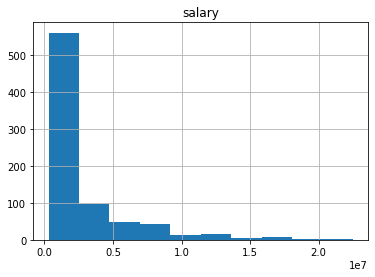
**Discussion Activity**

**2.1**

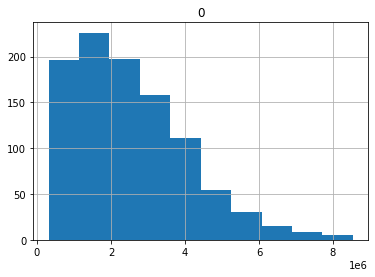
Dear Fellow Coworker,

I came across a dataset of Major League Baseball player salaries which was gathered from USA Today’s database. I thought I could see The *Central Limit Theorem* in action as work shown below.

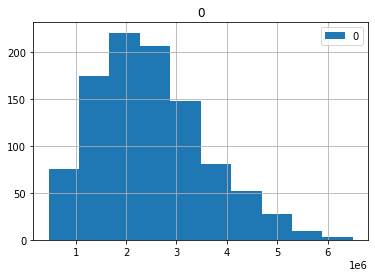
Please note that the dataset is right-skewed showing number of low earners are majority and there is a long tail towards high earners as shown in the diagram also reflected in the median and mean. The range of salaries from $300000 through $22500000, the median is $800000 and the mean is $2497668.6850690087, not linear, the standard deviation is 3535924.969930462.



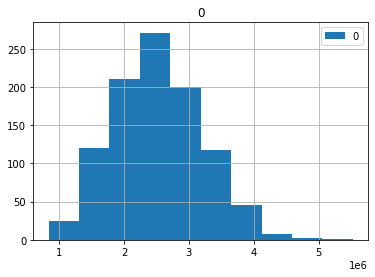
First, I started with 5 sample size, the mean is 2522718.052, it is close to the dataset’s mean value! But, the sample data shape is still skewed, the standard deviation is 1528540.9580613254 smaller than the dataset’s standard deviation value.



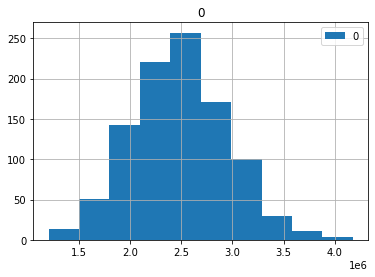
When the sample size is 10, the histogram becomes looking normal distribution, mean is more close to the original dataset’s mean value, 2481489.6061000004 now, the standard deviation is smaller 1089048.6400678856.



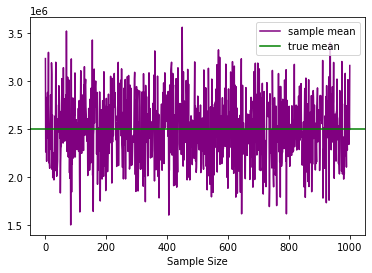
Starting with sample size 25, the histogram started looking better, however, the mean is a little off at 2516991.16008 but the standard deviation is more smaller 680793.4292964822.



However, at sample size 50, the histogram getting better, however, the mean is more closer to the original value at 2488753.0537199997, the standard deviation is more smaller 465761.1101933828.



At sample size 100, the mean at 2500692.7811100003 converges to the original dataset’s value as the *Central Limit Theorem* dictates the mean of a sample of data will be closer to the mean of the overall population as the sample size increases, see below the mean comparison chart, also note that the standard deviation is at its smallest 335622.5646845928.



So, as the histogram approaches to normal distribution, the standard deviation gets smaller as an indicator.

I just wanted to share my observations with you.

Regards,

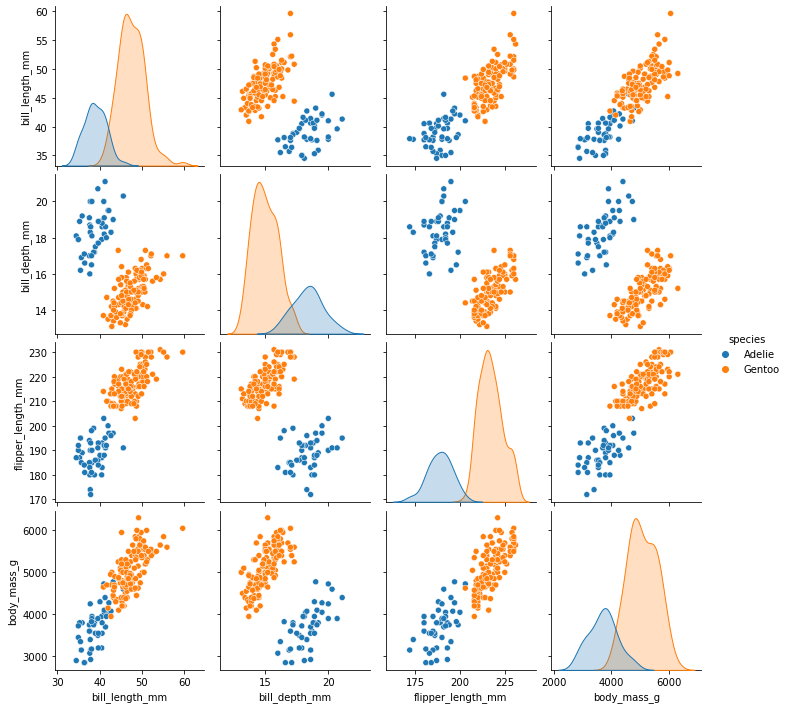
Aykan Erdenizmenli

**2.2**

**Penguins:**

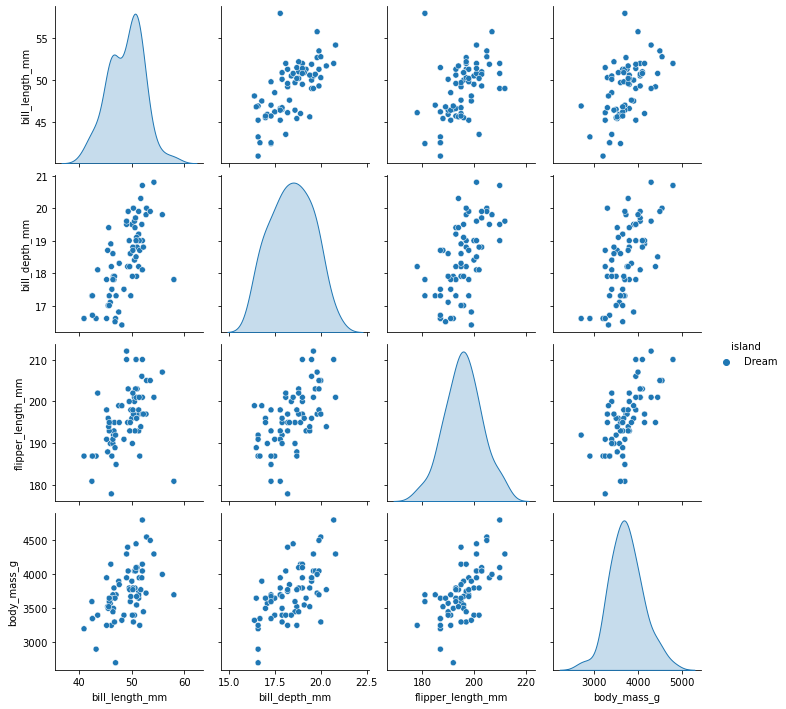
In penguins dataset, besides species there is also island information, I thought I could split data by islands to see if geography has a role. When I drew pair plot it is clear that penguins leave on Biscoe Island has larger range on all anatomical measurements. So, I could split data by Biscoe and other islands to start with.

There is strong positive correlation between body mass with all other measurements by each species which are Adelie and Gentoo and every other measurement as shown below.



In contrast to that there is weak positive correlation in the other islands dataset between those numeric features even for Adelie specie on Torgersen and Dream islands.

Next step was to separating **Chinstrap** from this other island dataset to analyze further: There is weak positive correlation between bill length with bill depth, flipper length, body mass; and between all other numeric measurements for this specie.



**Cars:**

What is axle ratio? It's a ratio that represents the number of revolutions the driveshaft must make to spin the axle one full turn. The higher the ratio, the higher the performance. The lower the ratio, the higher the fuel economy.

**First set by categorical feature of engine type (vs):**

Negative correlation mpg with engine size, higher the engine size less mileage per gallon.

Negative correlation mpg with hp, more the horse power is less mileage per gallon.

Positive correlation mpg with rear axle ratio, the higher the ratio is more mileage per gallon.

Negative correlation mpg with weight, more the weight is less mileage per gallon.

No correlation mpg with 1/4 mile time.

Positive correlation engine size with horse power.

Negative correlation engine size with rear axle ratio, higher the engine size is smaller the ratio.

Positive correlation engine size with weight.

Negative correlation engine size with 1/4 mile time.

Negative correlation horse power with rear axle ratio, higher the horse power is smaller the ratio.

Positive correlation horse power with weight.

Negative correlation horse power with 1/4 mile time.

Negative correlation rear axle ratio with weight, higher the ration is lighter weight

Positive correlation ear axle ratio with 1/4 mile time.

No correlation weight with 1/4 mile time.

**Second set by categorical feature of manual/automatic transmission (am):**

Negative correlation mpg with engine size, higher the engine size less mileage per gallon.

Negative correlation mpg with hp, more the horse power is less mileage per gallon.

Positive correlation mpg with rear axle ratio, the higher the ratio is more mileage per gallon.

Negative correlation mpg with weight, more the weight is less mileage per gallon.

No correlation mpg with 1/4 mile time.

Positive correlation engine size with horse power.

Negative correlation engine size with rear axle ratio, higher the engine size is smaller the ratio.

Positive correlation engine size with weight.

Negative correlation engine size with 1/4 mile time.

Negative correlation horse power with rear axle ratio, higher the horse power is smaller the ratio.

Positive correlation horse power with weight.

Negative correlation horse power with 1/4 mile time.

Negative correlation rear axle ratio with weight, higher the ration is lighter weight

Positive correlation ear axle ratio with 1/4 mile time.

No correlation weight with 1/4 mile time.

**Third set by categorical feature of number of forward gears (gear):**

Negative correlation mpg with engine size, higher the engine size less mileage per gallon.

Negative correlation mpg with hp, more the horse power is less mileage per gallon.

Positive correlation mpg with rear axle ratio, the higher the ratio is more mileage per gallon.

Negative correlation mpg with weight, more the weight is less mileage per gallon.

No correlation mpg with 1/4 mile time.

Positive correlation engine size with horse power.

Negative correlation engine size with rear axle ratio, higher the engine size is smaller the ratio.

Positive correlation engine size with weight.

Negative correlation engine size with 1/4 mile time.

Negative correlation horse power with rear axle ratio, higher the horse power is smaller the ratio.

Positive correlation horse power with weight.

Negative correlation horse power with 1/4 mile time.

Negative correlation rear axle ratio with weight, higher the ratio is lighter weight

Positive correlation ear axle ratio with 1/4 mile time.

No correlation weight with 1/4 mile time.

**Fourth set by categorical feature of number of carburetors (carb): (except 3, 6 and 8 carburetors)**

Negative correlation mpg with engine size, higher the engine size less mileage per gallon.

Negative correlation mpg with hp, more the horse power is less mileage per gallon.

Positive correlation mpg with rear axle ratio, the higher the ratio is more mileage per gallon.

Negative correlation mpg with weight, more the weight is less mileage per gallon.

No correlation mpg with 1/4 mile time.

Positive correlation engine size with horse power.

Negative correlation engine size with rear axle ratio, higher the engine size is smaller the ratio.

Positive correlation engine size with weight.

Negative correlation engine size with 1/4 mile time.

Negative correlation horse power with rear axle ratio, higher the horse power is smaller the ratio.

Positive correlation horse power with weight.

Negative correlation horse power with 1/4 mile time.

Negative correlation rear axle ratio with weight, higher the ratio is lighter weight

Positive correlation ear axle ratio with 1/4 mile time.

No correlation weight with 1/4 mile time.

**Fifth set by categorical feature of number of cylinders (cyl):**

Negative correlation mpg with engine size, higher the engine size less mileage per gallon.

Negative correlation mpg with hp, more the horse power is less mileage per gallon.

Positive correlation mpg with rear axle ratio, the higher the ratio is more mileage per gallon.

Negative correlation mpg with weight, more the weight is less mileage per gallon.

No correlation mpg with 1/4 mile time.

Positive correlation engine size with horse power.

Negative correlation engine size with rear axle ratio, higher the engine size is smaller the ratio.

Positive correlation engine size with weight.

Negative correlation engine size with 1/4 mile time.

Negative correlation horse power with rear axle ratio, higher the horse power is smaller the ratio.

Positive correlation horse power with weight.

Negative correlation horse power with 1/4 mile time.

Negative correlation rear axle ratio with weight, higher the ratio is lighter weight

Positive correlation ear axle ratio with 1/4 mile time.

No correlation weight with 1/4 mile time.

——

**Cars:**

**MPG Correlation with:**

Negative: engine size, hp, weight

Positive: rear axle ratio

None with 1/4 mile time.

**Engine Size Correlation with:**

Positive: horse power, weight.

Negative: rear axle ratio, 1/4 mile time.

**Horse Power Correlation:**

Negative: rear axle ratio, 1/4 mile time.

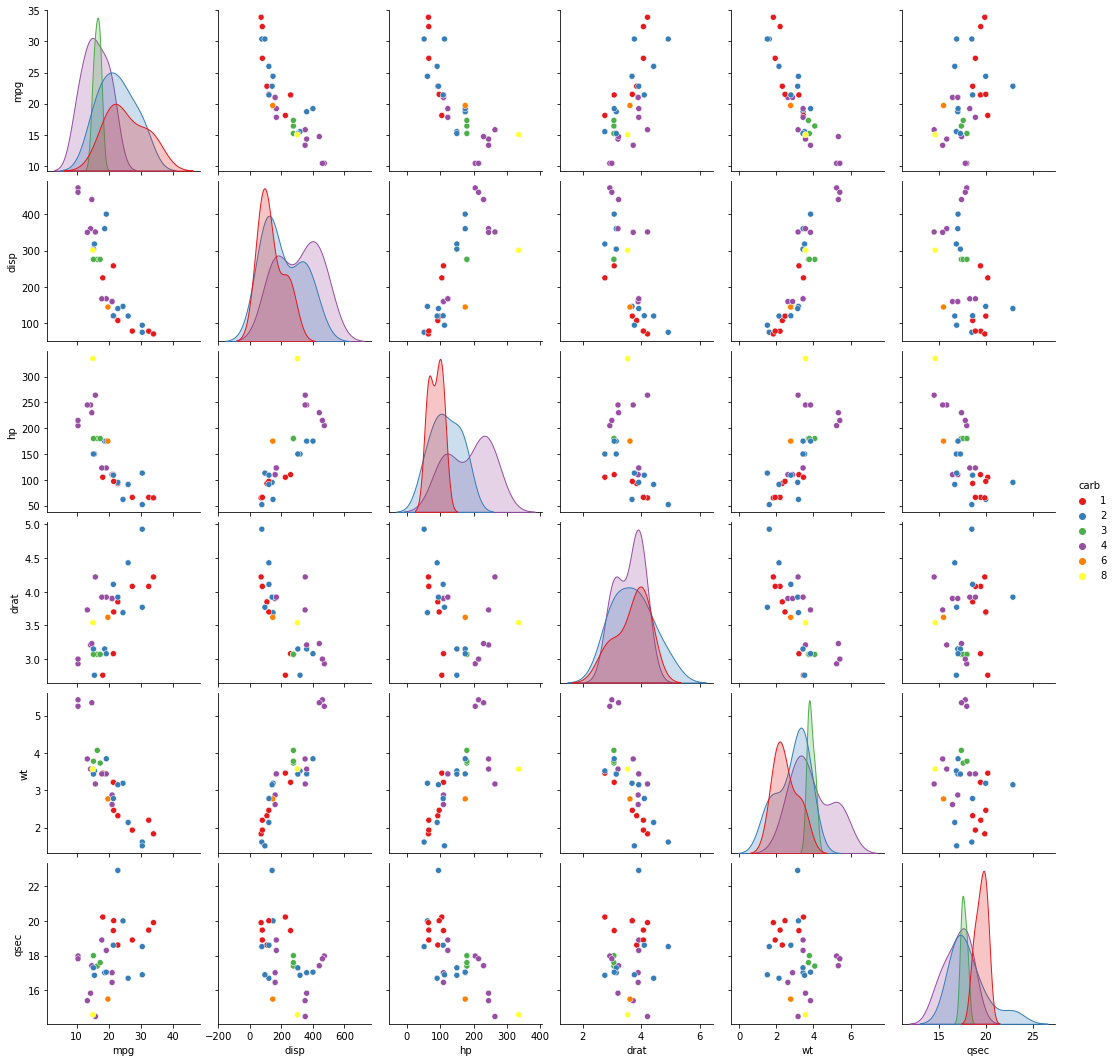
Positive: weight.

**Rear Axle Ratio Correlation:**

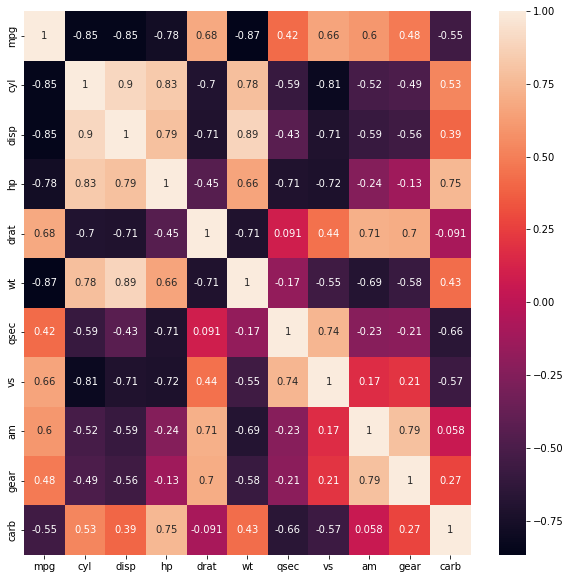
Negative: weight.

Positive: 1/4 mile time.

**No correlation weight with 1/4 mile time.**



All 5 datasets go with same correlation except (carb) categorical feature of 3, 6 and 8 as there are not many data points to conclude!



 ————— o —————