Assignment 2

Karla Aniela Cepeda Zapata

D00242569

Hi Siobhán:

Please, find the solutions from Assignment 2, Section A.
Also, you will find the .py file in Code folder. I tried to comment the code as much as possible to make clear what exactly I did.

Regards,
Karla Cepeda

Assignment 2 (10%):
Due at 23:59 on the 22th of November

For Section A provide all the code (.py file) and report the answers with results and graphs from python in a document along with your interpretation. Section B: Answer using pen and paper, Scan/Photograph. Submit to Moodle by zipping the folder

Section A: Cats' Heart Weight

Load in the data "Cat_Hwt.csv" from the dataset available in moodle. Data was collected on male and female adult cats used for experiments:
Sex: "F" for female and "M" for male.
Bwt: body weight in kg.
Hwt: heart weight in g.
Height: Height in cm
Age: Age in years.

(a) Describe the dataset, dimensions and what type of variables there are.

Description of the dataset
• *The dataset is a sample of adult cats for experiments.* • *Looking at the dataset, it is made up of 5 columns and 144 rows.* • *The function .info() says the data in the first column is an object type (probably string).* • *Also, it is stated that all rows and columns have non-null values, i.e. we have no missing values.*

Types of variables
• *Sex: string, one character. ('F','M') where 'F' stands for 'female' and 'M' for male.* • *Body weight: decima in kg.* • *Heart weight: decima in g.* • *Age: decimal, in years.*

Dimensions and type of variables	
Sex	***Categorical Nominal, (Dtype object)***
Bwt *(i.e. body weight in kg)*	***Numerical Continuous (Dtype float64)***
Hwt *(i.e. heart weight in g)*	***Numerical Continuous (Dtype float64)***
Height	***Numerical Continuous (Dtype float64)***
Age	***Numerical Continuous (Dtype float64)***

Additionally, we have 144 observations (i.e. 144 adult cats).
In total, we have 5*144 = 720 data.

```
In [155]: cats.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 144 entries, 0 to 143
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Sex     144 non-null    object
 1   Bwt     144 non-null    float64
 2   Hwt     144 non-null    float64
 3   Height  144 non-null    float64
 4   Age     144 non-null    float64
dtypes: float64(4), object(1)
memory usage: 5.8+ KB
```

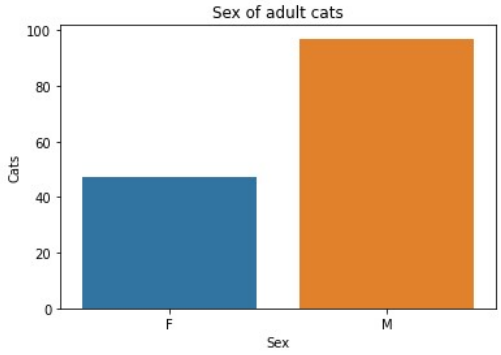
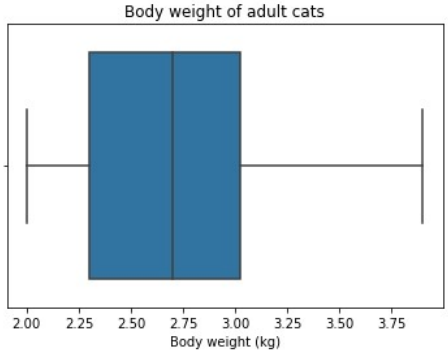
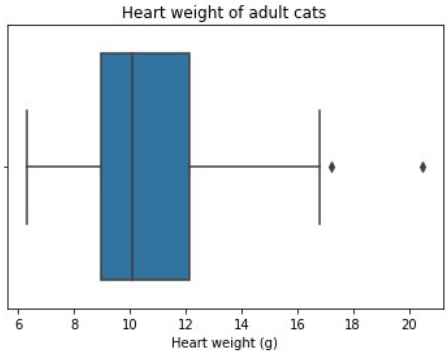
(b) Is there any missing data or outliers? If so, how do you recommend proceeding?

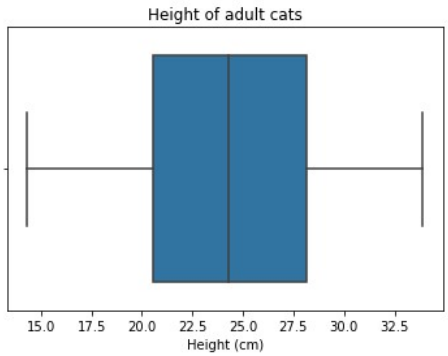
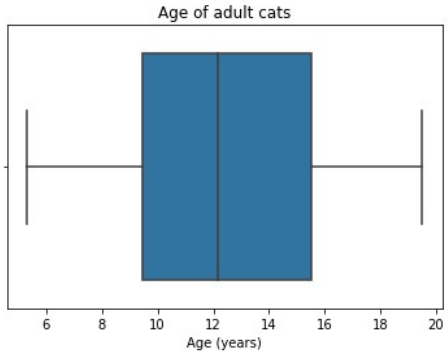
The function .info() from Python indicates that all the 144 observations in each variable are non-null. In other words, there are no missing values.
There are two outliers in the column Hwt: 17.2 g and 20.5 g:

- ***First of all, it does not see like the data has been taken wrongly.***
- ***Since maybe these male cats could be overweight, taking in consideration that the weight of the heart is related to the body (just as an assumption), I decided to get the ratio hwt:bwt and see if there is an outlier. There is no outlier. Actually, by plotting a boxplot the data seems pretty symmetric.***
- ***I am going to leave the outliers for this question.***

```
In [154]: boxplot_stats(cats.Hwt)[0]['fliers'] # Two outliers, 17.2 and 20.5
Out[154]: array([17.2, 20.5])
```

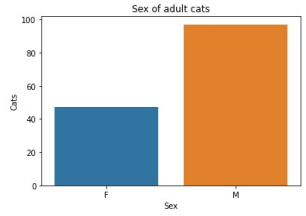
(c) Create univariate plots and interpret the plots.

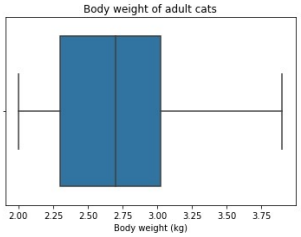
Graph	***Interpretation***
	• ***The graph shows that there are more males than females in the sample of adult cats.***
	• ***There is a slight concentration of data in Q3 compared to Q2.*** • ***50% of data is balanced to the right.*** • ***Therefore, positive skewness is shown in the histogram. Skewed-right distribution.*** • ***No outliers***
	• ***There is a strong concentration of data in Q2 compared to Q3.*** • ***50% of data is slightly balanced to the right.*** • ***Therefore, positive right skewness is shown.*** • ***There are two outliers.***

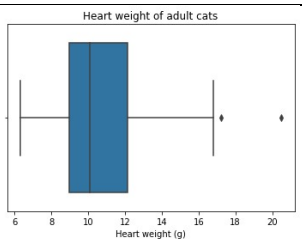
	• ***Q2 and Q3 data is symmetric. Symmetric box.*** • ***Whiskers have same length.*** • ***Therefore, plot shown a bell shape. No skewed data.*** • ***No outliers.***
	• ***Q2 has slightly more data concentrated comparing to Q3.*** • ***Whiskers have slightly same length.*** • ***Data is ok. Therefore, plot shown a symmetric shape.*** • ***No outliers.***

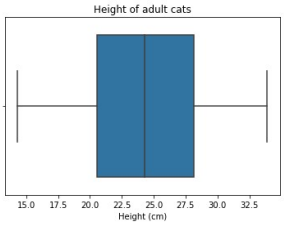
(d) Based on the plots, examine appropriate summary statistics and interpret

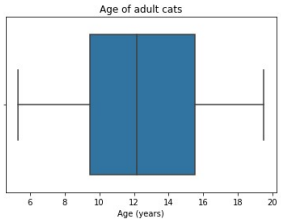
Graph and summary statistics	*Interpretation*


`In [147]: cats['Sex'].value_counts()`
`Out[147]:`
`M  97`
`F  47`
`Name: Sex, dtype: int64`
	• ***There are 47 female cats and 97 male cats, which summing up give 144 observations.***


`In [148]: cats.Hwt.describe()`
`Out[148]:`
`count  144.000000`
`mean  10.630556`
`std  2.434636`
`min  6.300000`
`25%  8.950000`
`50%  10.100000`
`75%  12.125000`
`max  20.500000`
`Name: Hwt, dtype: float64`
	• ***The lowest body weight among the sample of adult cats is 2.0 kg***
	• ***The highest body weight among the sample of adult cats is 3.9 kg***
	• ***The average body weight in the sample is 2.72 kg,***
	• ***The median body weight is 2.70 kg***
	• ***NOTE: it is interesting that mean and median has almost same value despite skewness.***


`In [149]: cats.Hwt.describe()`
`Out[149]:`
`count  144.000000`
`mean  10.630556`
`std  2.434636`
`min  6.300000`
`25%  8.950000`
`50%  10.100000`
`75%  12.125000`
`max  20.500000`
`Name: Hwt, dtype: float64`
	• ***The lowest heart weight in the sample is 6.3 g***
	• ***The highest heart weight in the sample is 20.5 g***
	• ***The average heart weight among adult cats in sample is 10.6 g,***
	• ***The median heart weight in the sample is 10.1 g***


`In [150]: cats.Height.describe()`
`Out[150]:`
`count  144.000000`
`mean  24.318056`
`std  5.156307`
`min  14.300000`
`25%  20.550000`
`50%  24.300000`
`75%  28.150000`
`max  33.900000`
`Name: Height, dtype: float64`
	• ***The largest cat is 33.9 cm***
	• ***The shortest cat is 14.3 cm***
	• ***The average height in the sample is 24.3 cm,***
	• ***The median height of adult cats is 24.3 cm***


`In [151]: cats.Age.describe()`
`Out[151]:`
`count  144.000000`
`mean  12.210417`
`std  3.850687`
`min  5.300000`
`25%  9.475000`
`50%  12.150000`
`75%  15.500000`
`max  19.500000`
`Name: Age, dtype: float64`
	• ***The oldest cat is 19.5 years***
	• ***The youngest cat is 5.3 years***
	• ***The average age in the sample is 12.2 years***
	• ***The median age of adult cats is 12.1 years***

(e) Create bivariate plots to explore the relationship between all pairs of variables. Interpret each plot.

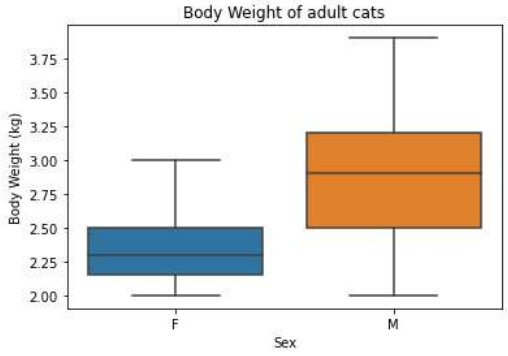
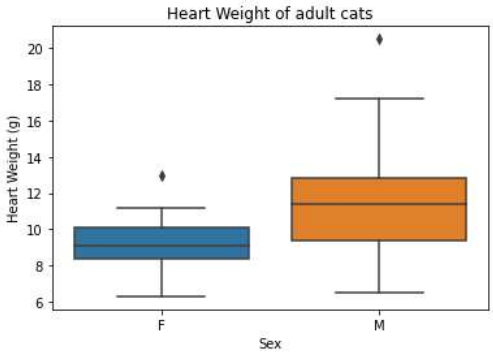
Let's create all pairs of variables:

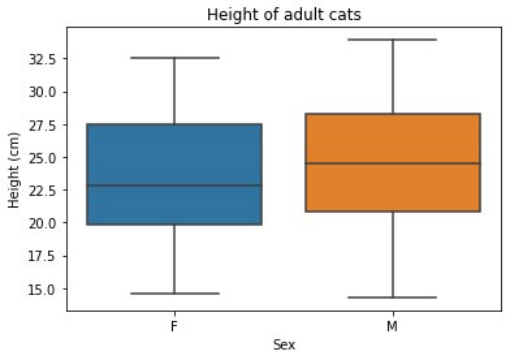
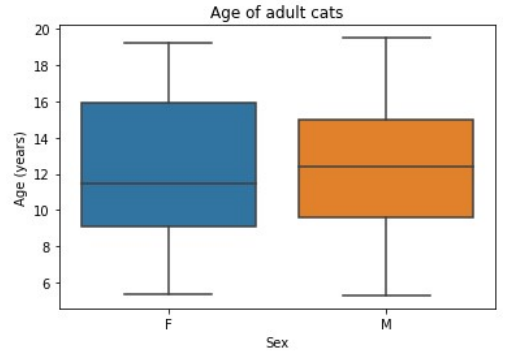
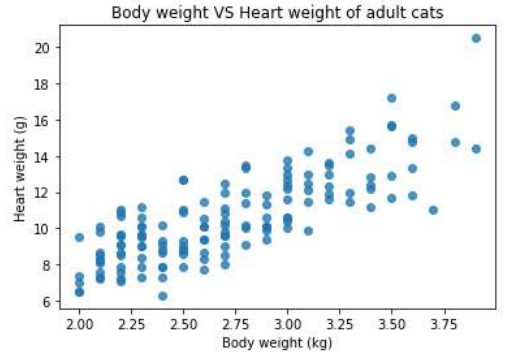
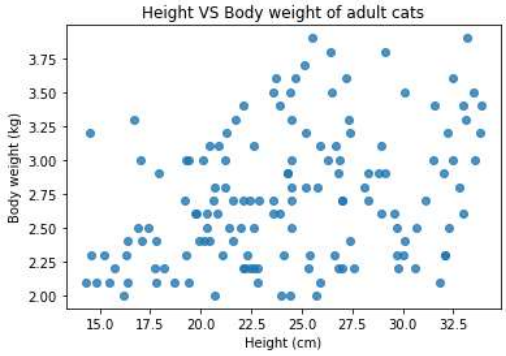
- ***Sex, Bwt => boxplot***
- ***Sex, Hwt => boxplot***
- ***Sex, Height => boxplot***
- ***Sex, Age => boxplot***

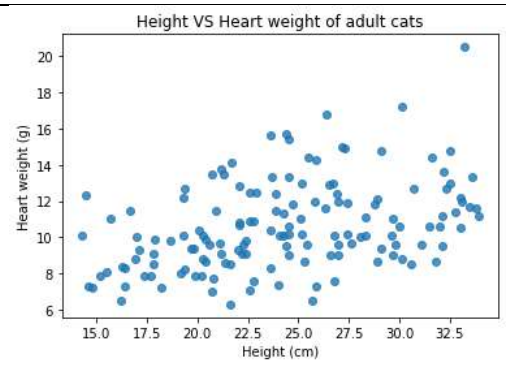
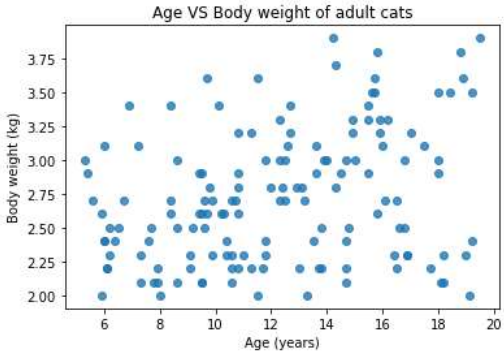
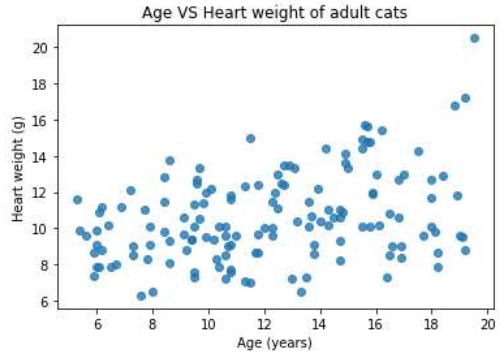
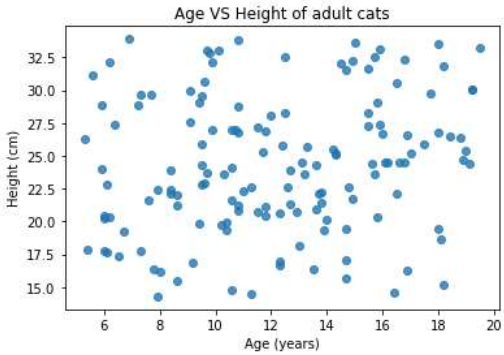
No more combinations since Bwt, Hwt, Height and Age are numerical continuous variables.

Next, for all numerical continuous variables, taking that x=independent, y=dependent for scatterbox:

- ***x=Bwt, y=Hwt => scatterbox***
- ***x=Height, y=Bwt => scatterbox***
- ***x=Height, y=Hwt => scatterbox***
- ***x=Age, y=Bwt => scatterbox***
- ***x=Age, y=Hwt => scatterbox***
- ***x=Age, y=Height => scatterbox***

Graph and summary statistics	***Interpretation***
	• ***There is a difference between medians.*** • ***The median is noticeable higher for male cats.*** • ***IQR from female cats do not overlaps.*** • ***Spread is not the same in both groups.*** • ***For group F: skewed shape.*** • ***For group M: symmetric shape.*** • ***Both groups have one outlier each.***
	• ***There is a difference between medians.*** • ***The median is smaller among female cats.*** • ***IQR from female cats overlaps just half.*** • ***Spread is not the same in both groups.*** • ***For group F: symmetric shape.*** • ***For group M: symmetric shape.*** • ***Both groups have one outlier each.***

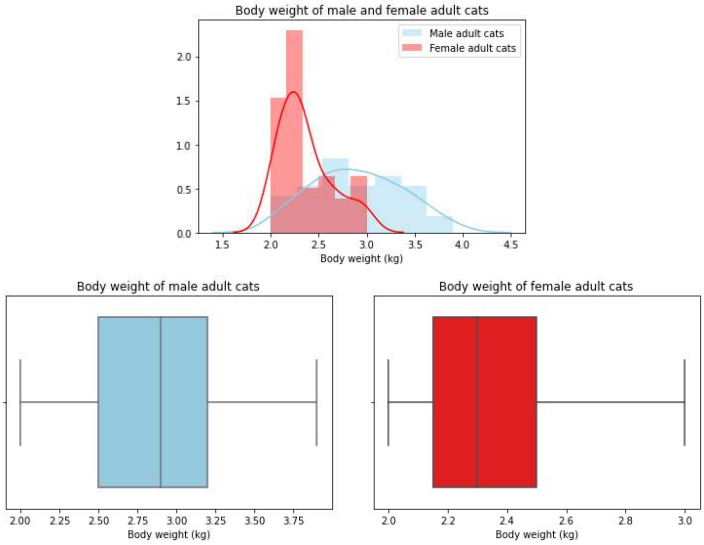
	• ***The lowest heart weight in the sample is 6.3 g*** • ***The highest heart weight in the sample is 20.5 g*** • ***The average heart weight among adult cats in sample is 10.6 g,*** • ***The median heart weight in the sample is 10.1 g***
	• ***There is a slight difference between medians, but male cats' IQR overlaps*** • ***The median is slightly higher among male cats.*** • ***IQR from male cats overlaps the IQR from female cats' group.*** • ***Spread is the same for both groups.*** • ***For group F: roughly symmetric shape.*** • ***For group M: symmetric shape.*** • ***No outliers.***
	• ***There is a positive relationship between body weight and heart weight***
	• ***There is a no relationship between height and body weight among adult cats*** • ***There is a random scatter***

	• ***There is a weak positive lineal relationship between height and heart weight***
	• ***There is no relationship between age and body weight among adult cats*** • ***There is a random scatter***
	• ***There is a weak positive lineal relationship between age and body weight.***
	• ***There is no relationship between age and height among cats.*** • ***There is a random scatter***

(f) Using the bivariate plots that appear to have a difference between two groups only, determine if there is a statistical difference between the groups using hypothesis testing. Make sure in your answer to explain the hypotheses, any assumptions needed and if they are met, results and interpretation of all the results. Conclude your findings.

For all the following tests, significant level value will be taken as 5%. i.e. $\alpha = 0.05$

For Body Weight VS Sex
Previously, in the bivariable boxplot there was shown a remarkable difference between medians.



Assumptions for two-samples t-test:

- ***Continuous data. Yes.***
- ***Samples must be independent and random. Yes.***
- ***Not be skewed (i.e. it has normally distributed / bell shaped). No, sample from females is strong skewed to the right and male is slightly skewed to the right.***
- ***Standard deviation must be the same (i.e. spread should be roughly the same). No.***

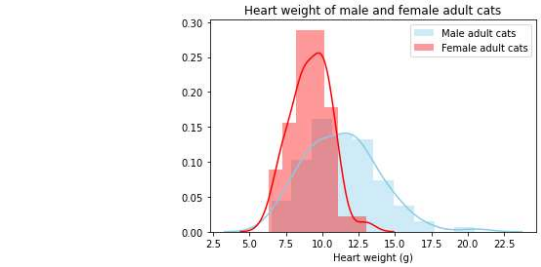
T-test cannot be performed since violation in assumptions enlisted above. Therefore, a non-parametric test would be a better option. Boxplots show that 50% of the data is slightly balanced to the right. In this case, a Wilcoxon-Mann-Whitney test is enough to perform a hypothesis test.

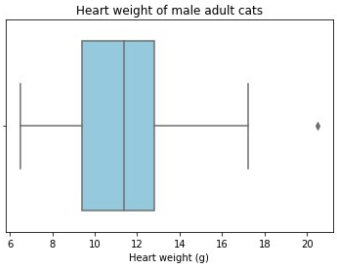
Hypothesis:
H_0: ***median body weight of female cats*** $=$ ***median body weight of male cats***
H_1: ***median body weight of female cats*** $!=$ ***median body weight of female cats***

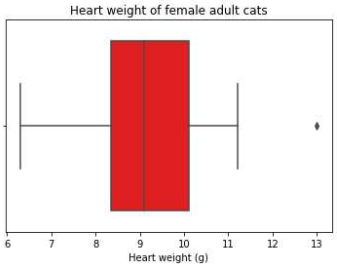
```
In [15]: m_cats_b=cats[cats["Sex"] == 'M']['Bwt']
    ...: f_cats_b=cats[cats["Sex"] == 'F']['Bwt']
    ...:
    ...: sns.boxplot(m_cats_b, color="skyblue")
    ...: plt.title("Body weight of male adult cats")
    ...: plt.xlabel("Body weight (kg)")
    ...:
    ...: sns.boxplot(f_cats_b,color="red")
    ...: plt.title("Body weight of female adult cats")
    ...: plt.xlabel("Body weight (kg)")
    ...:
    ...: sns.distplot(m_cats_b, color="skyblue" ,label="Male adult cats")
    ...: sns.distplot(f_cats_b, color="red", label="Female adult cats")
    ...: plt.title("Body weight of male and female adult cats")
    ...: plt.xlabel("Body weight (kg)")
    ...: plt.legend()
    ...: ranksums(m_cats_b,f_cats_b)
Out[15]: RanksumsResult(statistic=6.484649068728496,
pvalue=8.893855174450822e-11)
```

$P - Value = 1.64 \times 10^{-10} < 0.5$. ***Therefore, we reject null hypothesis, there is a difference between medians. In other words, there is a relationship between sex and body weight. There is not enough evidence to suggest that body weight is not affected by sex.***

For Sex - Heart weight:
Previously, in the bivariable boxplot there was shown a slight difference between medians.







Assumptions for two-samples t-test:

- ***Continuous data. Yes.***
- ***Samples must be independent and random. Yes.***
- ***Not be skewed (i.e. it has normally distributed / bell shaped). Yes, roughly symmetric shape.***
- ***Standard deviation must be the same (i.e. spread should be roughly the same). No.***

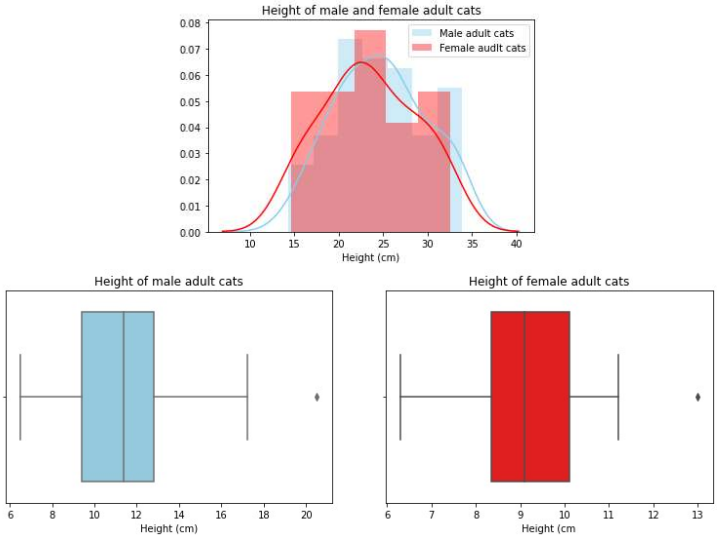
T-test cannot be performed since violation in assumptions enlisted above. Therefore, a non-parametric test would be the best option. Wilcoxon-Mann-Whitney test would be performed.

Hypothesis:
H_0: ***median heart weight of female cats*** $=$ ***median heart weight of male cats***
H_1: ***median heart weight of female cats*** $!=$ ***median heart weight of female cats***

```
In [21]: m_cats_h=cats[cats["Sex"] == 'M']['Hwt']
    ...: f_cats_h=cats[cats["Sex"] == 'F']['Hwt']
    ...:
    ...: sns.boxplot(m_cats_h, color="skyblue")
    ...: plt.title("Heart weight of male adult cats")
    ...: plt.xlabel("Heart weight (g)")
    ...:
    ...: sns.boxplot(f_cats_h,color="red")
    ...: plt.title("Heart weight of female adult cats")
    ...: plt.xlabel("Heart weight (g)")
    ...:
    ...: sns.distplot(m_cats_h, color="skyblue" ,label="Male adult cats")
    ...: sns.distplot(f_cats_h, color="red", label="Female adult cats")
    ...: plt.title("Heart weight of male and female adult cats")
    ...: plt.xlabel("Heart weight (g)")
    ...: plt.legend()
    ...: ranksums(m_cats_h, f_cats_h)
Out[21]: RanksumsResult(statistic=5.031780913382624,
pvalue=4.859444084159132e-07)
```

$P - Value = 4.86 \times 10^{-7} < 0.05$. ***Therefore, we reject the null hypothesis, there is a difference between medians. In other words, there is a relationship between sex and heart weight. There is not enough evidence to suggest that heart weight is not affected by sex.***

For Height - Sex:
Previously, in the bivariable boxplot there was shown a slight difference between medians.



Assumptions for two-samples t-test:

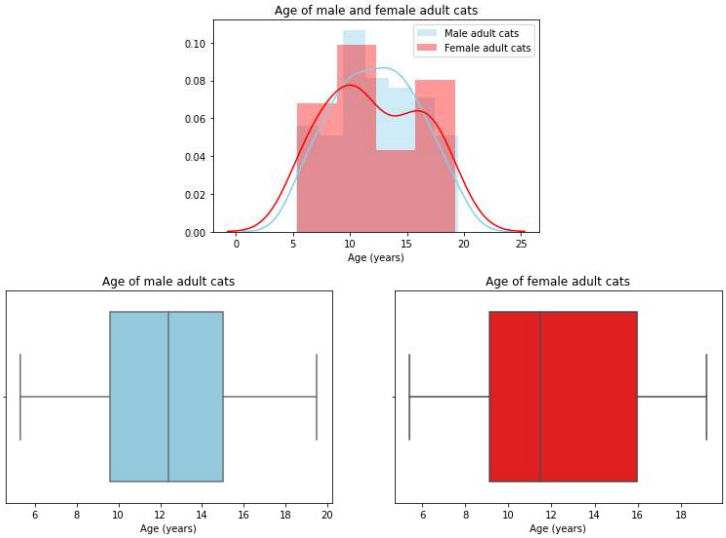
- ***Samples must be independent and random. Yes.***
- ***Not be skewed (i.e. it has normally distributed / bell shaped). Yes, it is roughly symmetric.***
- ***Standard deviation must be the same (i.e. spread should be roughly the same). Yes.***

Hypothesis:
H_0: $mean\ height\ of\ female\ cats = mean\ height\ of\ male\ cats$
H_1: $mean\ height\ of\ female\ cats\ ! = mean\ height\ of\ male\ cats$

```
In [24]: m_cats_he=cats[cats["Sex"] == 'M']['Height']
    ...: f_cats_he=cats[cats["Sex"] == 'F']['Height']
    ...:
    ...: sns.distplot(m_cats_he, color="skyblue" ,label="Male adult cats")
    ...: sns.distplot(f_cats_he, color="red", label="Female audlt cats")
    ...: plt.title("Height of male and female adult cats")
    ...: plt.xlabel("Height (cm)")
    ...: plt.legend()
    ...:
    ...: ttest_ind(m_cats_he, f_cats_he, equal_var=True)
Out[24]: Ttest_indResult(statistic=1.4023174744446743,
pvalue=0.16300282308176756)
```

$P - Value = 0.16 > 0.05$. ***Therefore, we fail to reject the null hypothesis, there is no difference between means. In other words, there is no relationship between sex and height. There is not enough evidence to suggest that height is affected by sex.***

For Age among Sex:
Previously, in the bivariable boxplot there was shown a slight difference between medians.



Assumptions for two-samples t-test:

- **Samples must be independent and random. Yes.**
- **Not be skewed (i.e. it has normally distributed / bell shaped). Yes, roughly symmetric.**
- **Standard deviation must be the same (i.e. spread should be roughly the same). Yes.**

Hypothesis:
H_0: $mean\ age\ of\ female\ cats = mean\ age\ of\ male\ cats$
H_1: $mean\ age\ of\ female\ cats\ ! = mean\ age\ of\ male\ cats$

```
In [31]: m_cats_a=cats[cats["Sex"] == 'M']['Age']
    ...: f_cats_a=cats[cats["Sex"] == 'F']['Age']
    ...:
    ...: sns.boxplot(m_cats_a,color="skyblue")
    ...: plt.title("Age of male adult cats")
    ...: plt.xlabel("Age (years)")
    ...:
    ...: sns.boxplot(f_cats_a,color="red")
    ...: plt.title("Age of female adult cats")
    ...: plt.xlabel("Age (years)")
    ...:
    ...: sns.distplot(m_cats_a, color="skyblue" ,label="Male adult cats")
    ...: sns.distplot(f_cats_a, color="red", label="Female adult cats")
    ...: plt.title("Age of male and female adult cats")
    ...: plt.xlabel("Age (years)")
    ...: plt.legend()
    ...:
    ...: ttest_ind(m_cats_a, f_cats_a, equal_var=True)
Out[31]: Ttest_indResult(statistic=0.22032120573768235,
pvalue=0.8259374476550417)
```

$P - Value = 0.83 > 0.05$. ***Therefore, we fail to reject the null hypothesis, there is no difference between means. In other words, there is no relationship between sex and age.***