FIT5145 - Introduction to Data Science

Assignment 1

The aim of this assignment is to investigate and visualise data using various data science tools. It will test your ability to:

1. read data files in Python and extract related data from those files;
2. wrangle and process data;
3. use various graphical and non-graphical tools to perform exploratory data analysis and visualisation;
4. use basic tools for managing and processing big data; and
5. communicate your findings in your report.

You will need to submit two files:

1. The Python code as a Jupyter notebook file that you wrote to analyse and plot the data.
2. A PDF of your Jupyter notebook file containing your answers (code, figures and answers to all the questions). Make sure to include screenshots/images of the graphs you generate in order to justify your answers to all the questions. Marks will be assigned to PDF reports based on their correctness and clarity. -- For example, higher marks will be given to PDF reports containing graphs with appropriately labelled axes.

IMPORTANT NOTE - Zip file submission will have a penalty of 10%. Do not submit the separate files requested above together in one Zip file. As indicated in the rubric, marks will be deducted for this because it adds significantly to the time it takes for the markers to open up and access your assignments given that there are many students in this class.

Tasks

There are two tasks that you need to complete for this assignment, Task A and Task B. You need to use Python to complete the tasks.

Task A - Who are Data Scientists? Data Scientist Demographics

'What does a does a Data Scientist look like?', *'What is Data Science exactly?'*, *'Is Python or R better to learn for beginners?'*, *'Do you have to have a degree in Computer Science to be a Data Scientist?'* and *'Do data scientists earn as much as I think?'*.

Anjul Bhambri, the Vice Preseident of big data prodcts at IBM says this

> *'A data scientist is somebody who is inquisitive, who can stare at data and spot trends. It's almost like a Renaissance individual who really wants to learn and bring change to an organisation.'*

In this course, you have learned that the diversity in definitions, skill sets, tools, applications and knowledge domains that make data science challenging to define precisely. By completing the following questions, we hope you can get a more precise understanding.

The Data

Kaggle is the home of analytics and predictive modelling competitions. Data Science enthusiast, beginners to professionals, compete to create the best predictive models using datasets uploaded both by individuals and companies looking for insights. Prizes can be as high as $3 million US. In late 2017 a survey of Kaggle users was conducted and received over 16,700 responses. The dataset was, of course, made public and many insights have emerged since. We have taken a portion of the data set and heavily modified the data. Both to clean the data, a significant component of data science and to ensure original assignment submission.

Your Job

The following notebook has been constructed to provide you with directions (blue), assessed questions (brown) and background information. Responses to both blue directions and brown questions are assessable.

You will be required to write your own code. Underneath direction boxes, there will be empty cells with the comment #Your code. Insert new cells under this cell if required.

To respond to questions you should double click on the cell beneath each question with the comment Answer.

Please note, your commenting and adherence to Python code standards will be marked. This notebook has been designed to give you a template for how we expect Python Notebooks to be submitted for assessment. If you require further information on Python standards, please visit https://www.python.org/dev/peps/pep-0008/ (https://www.python.org/dev/peps/pep-0008/) Do not change any of the directions or answer boxes, the order of questions, order of code entry cells or the name of the input files.

The Files

- *multipleChoiceResponses.csv * : Participants' answers to multiple choice questions. Each column contains the answers of one respondent to a specific question.
- *conversionRates.csv* : Currency conversion rates to USD.

** Your Information** Enter your information in the following cell. Please make sure you specify what version of python you are using as your tutor may not be using the same version and will adjust your code accordingly.

Student Information

Please enter your details here.

Name: Shweta Sharma

Student number: 29888956

Tutorial Day and Time: Monday, 10am - 12Pm

Tutor: Zhinoos Razavi Hesabi

Environment: Python 3.7 and Anaconda 5.2.0-py36_3

Table of contents

0. Load your libraries and files

1. ** Load your libraries and files**
This assesment will be conducted using pandas. You will also be required to create visualisations. We recomend Seaborn which is more visually appealing than matplotlib. However, you may choose either. For further information on Seaborn visit https://seaborn.pydata.org/ (https://seaborn.pydata.org/)

Hint: Remember to comment what each library does.

In [1]:

```
# Your code
import pandas as pd                 # for loading CSV file and reading the data
import seaborn as sb                # For visualisation and creating graphs
import numpy as np                  # for mathematical/statistical operations
import matplotlib.pyplot as plt     # For visualisation and creating graphs
%matplotlib inline
```

1. Demographic Analysis

So what does a data scientist look like?

Let's get a general understanding of the characteristics of the survey participants. Demographic overviews are a standard way to start an exploration of survey data. The types of participants can heavily affect the survey responses.

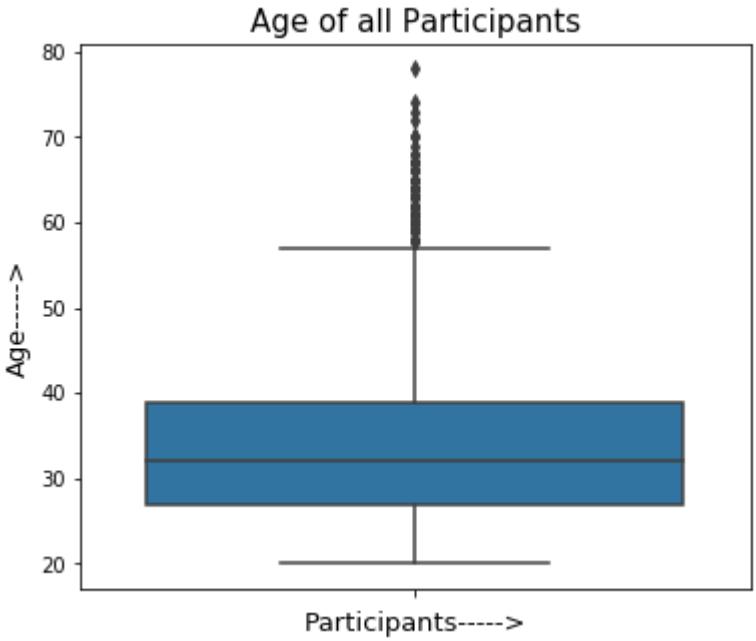
1.1 Age

Visualisation is a quick and easy way to gain an overview of the data. One method is through a boxplot. Boxplots are a way to show the distribution of numerical data and display the five descriptive statistics: minimum, first quartile, median, third quartile, and maximum. Outliers should also be shown.

2 Create a box plot showing the age of all the participants.
Your plot must have labels for each axis, a title, numerical points for the age axis and also show the outliers.

In [37]:

```
# Your code
import pandas as pd
response =pd.read_csv('multipleChoiceResponses.csv')
plt.figure(figsize=(6,5))
sb.boxplot(y=response['Age'])
plt.title('Age of all Participants',fontsize=15)
plt.ylabel('Age----->',fontsize=13)
plt.xlabel('Participants----->',fontsize=13)
plt.show()
```



3. Calculate the five descriptive statistics as shown on the boxplot, as well as the mean
Round your answer to the nearest whole number.

In [3]:

```
# Your code
#the five descriptive statistics: minimum, first quartile, median, third quartile,
minimum_age=round(min(response['Age']))
first_quartile,median_age,third_quartile=np.percentile(response['Age'],[25,50,75])
maximum_age=round(max(response['Age']))
Average_age=round(np.mean(response['Age']))

print("\nMinimum age: ",minimum_age,"\nFirst Quartile: ",round(first_quartile),"\nM
```

```
Minimum age:  20
First Quartile:  27.0
Median Age:  32.0
Third quartile 39.0
Maximum Age:  78
Mean of Age:  34
```

Answer The minimum age among the participants is : 20 Yrs The median of age of all the participants is : 32 Yrs The mean of age of all the participants is : 34 Yrs The first quantile(25%) of the age of participants is : 27 Yrs The third quantile(75%) of the age of participants is : 39 Yrs The maximum age among the participants is : 78 Yrs

4. Looking at the boxplot what general conclusion can you make about the age of the participants?
You must explain your answer concerning the median, minimum and maximum age of the respondents.
You must also make mention of the outliers if there are any.

Answer The box plot in Q2 clearly shows the descriptive statistics of the Age of participants. It show that most of the participants are young, between 27 to 39 years(between first quatile and third quantile), which determines why median is 32 years.Also, we have few outliers such as participant with age more than 65(the maximum limit of box plot), such as the participant with as of 78 years(also the maximum age).The yougest participant is of age 20 years.

5. Regardless of the errors that the data show, we are interested in working-age data scientists, aged between 18 and 65.
How many respondents were under 18 or over 65?

In [4]:

```
# Your code
#counting all the invalid participants
count=0
for each in response.Age:
    if ((each < 18)|(each > 65)):
        count=count+1

#counting the invalid participants who are working as Data Scientist
count1=0
# filterinf data for only Data Scientist
new_df=response[response.CurrentJobTitleSelect=='Data Scientist'].reset_index()
for each in new_df.Age:
    if ((each < 18)|(each > 65)):
        count1=count1+1

print("Number of respondents under 18 or over 65 years of age are : ",count)
print("Number of respondents with job title \"Data Scientist\" under 18 or over 65
```

```
Number of respondents under 18 or over 65 years of age are :  19
Number of respondents with job title "Data Scientist" under 18 or over
65 years of age are :  4
```

Answer The above code tell us that there are not many outliers, out of 3540 participants we have 19 participants who are either under 18 or over 65 years of age and in case of participants with job title as Data Scientist, there are only 4.

1.2 Gender

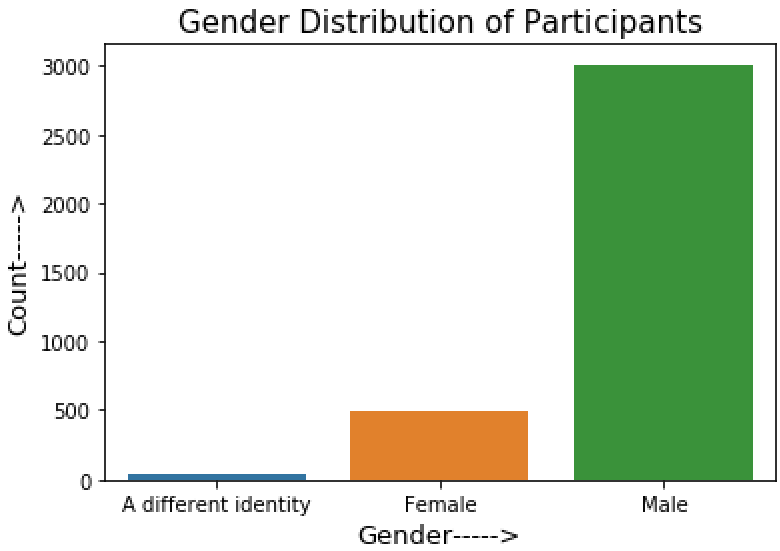
We are interested in the gender of respondents. Within the STEM fields, there are more males than females or other genders. In 2016 the Office of the chief scientist found that women held only 25% of jobs in STEM. Let's see how data science compares.

6. Plot the gender distribution of survey participants.

In [5]:

```
# Your code
# counting the number of participants for each gender type
fun = {'GenderSelect' : {'Respondant':'count'}}
gender_count=response.groupby('GenderSelect').agg(fun).reset_index()
gender_count.columns=gender_count.columns.droplevel(0)
gender_count.rename(columns={'':'Gender'},inplace =True )
sb.barplot(x=gender_count.Gender,y=gender_count.Respondant)
plt.title('Gender Distribution of Participants',fontsize=15)
plt.xlabel('Gender----->',fontsize=13)
plt.ylabel('Count----->',fontsize=13)
plt.show()
```

```
/home/mranali/anaconda3/lib/python3.7/site-packages/pandas/core/groupb
y/groupby.py:4656: FutureWarning: using a dict with renaming is deprec
ated and will be removed in a future version
  return super(DataFrameGroupBy, self).aggregate(arg, *args, **kwargs)
```



7. What percentage of respondants were men? What percentage of respondants were women?

In [6]:

```
# Your code
total_respondant=sum(gender_count.Respondant)
n=len(gender_count.Gender)
for each in range(n):
    if gender_count.Gender[each] == 'Male':
        # calculating percentage of male respondantes round to two decimal places
        male_percentage= round((gender_count.Respondant[each] /total_respondant)*10
        # calculating percentage of female respondantes round to two decimal places
    if gender_count.Gender[each] == 'Female':
        female_percentage= round((gender_count.Respondant[each] /total_respondant)*

print("Percentage of male respondants :",male_percentage)
print("Percentage of female respondants :",female_percentage)
```

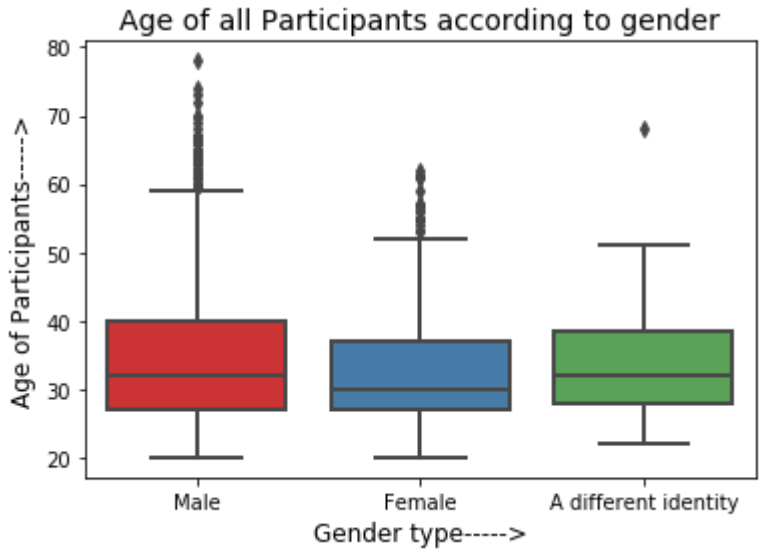
```
Percentage of male respondants : 84.97
Percentage of female respondants : 14.01
```

Answer 84.97% of participants are Men and 14.01%(rounded to two decimal places) are Women.

8. Let's see if there is any relationship between age and gender.
Create a box plot showing the age of all the participants according to gender.
Include the response 'Different identity' in your plot.

In [7]:

```
# Your code
sb.boxplot(x='GenderSelect',y='Age',data=response,palette="Set1",linewidth=2)
plt.title('Age of all Participants according to gender',fontsize=14)
plt.xlabel('Gender type----->',fontsize=12)
plt.ylabel('Age of Participants----->',fontsize=12)
plt.show()
```



9. What comments can you make about the relationship between the age and gender of the respondents?

Hint: You need to determine the numeric descriptive statistics

In [8]:

```
# Your code
n=len(response.GenderSelect) #number of participants
male_age=[]
female_age=[]
others_age=[]
for i in range(n):
    if response.GenderSelect[i]=='Male':
        male_age.append(response.Age[i])
    elif response.GenderSelect[i]=='Female':
        female_age.append(response.Age[i])
    else:
        others_age.append(response.Age[i])

# Definig function to calculate the numeric descriptions
def statsistics(List):
    minimum_age=np.min(List)
    maximum_age=np.max(List)
    meadian_age=np.median(List)
    mean_age=np.mean(List)
    first_quartile=np.percentile(List,25)
    third_quartile=np.percentile(List,75)
    statsistics=[minimum_age,maximum_age,meadian_age,mean_age,first_quartile,third_
    return statsistics

Male_stats=statsistics(male_age)
Female_stats=statsistics(female_age)
Diff_identity_stats=statsistics(others_age)
df = pd.DataFrame({'Statistics':['Minimum_age','Maximum_age','Meadian_age','Mean_ag
df
```

Out[8]:

	Statistics	Male	Female	Different Identity
0	Minimum_age	20.000000	20.000000	22.000000
1	Maximum_age	78.000000	62.000000	68.000000
2	Meadian_age	32.000000	30.000000	32.000000
3	Mean_age	34.637633	32.735887	34.666667
4	First_quartile	27.000000	27.000000	28.000000
5	Third_quartile	40.000000	37.000000	38.500000

Answer From the above statistics we can see that the average age of working class of all the three gender is approximately same 35 years for Men and Different Identity and 33 Years for Women.But as the age Increases, women tend to quit(guessing), we see by third quartile we have age gap of 3 years between Women(37 years)

and Men(40 years).The gap is even more when it comes to maximum working age, for Men it is 78 years where as for women it is 62 years.Participants with Different Identity are better than Women with maximum working age of 68 Years, but still behind Men.

1.3 Country

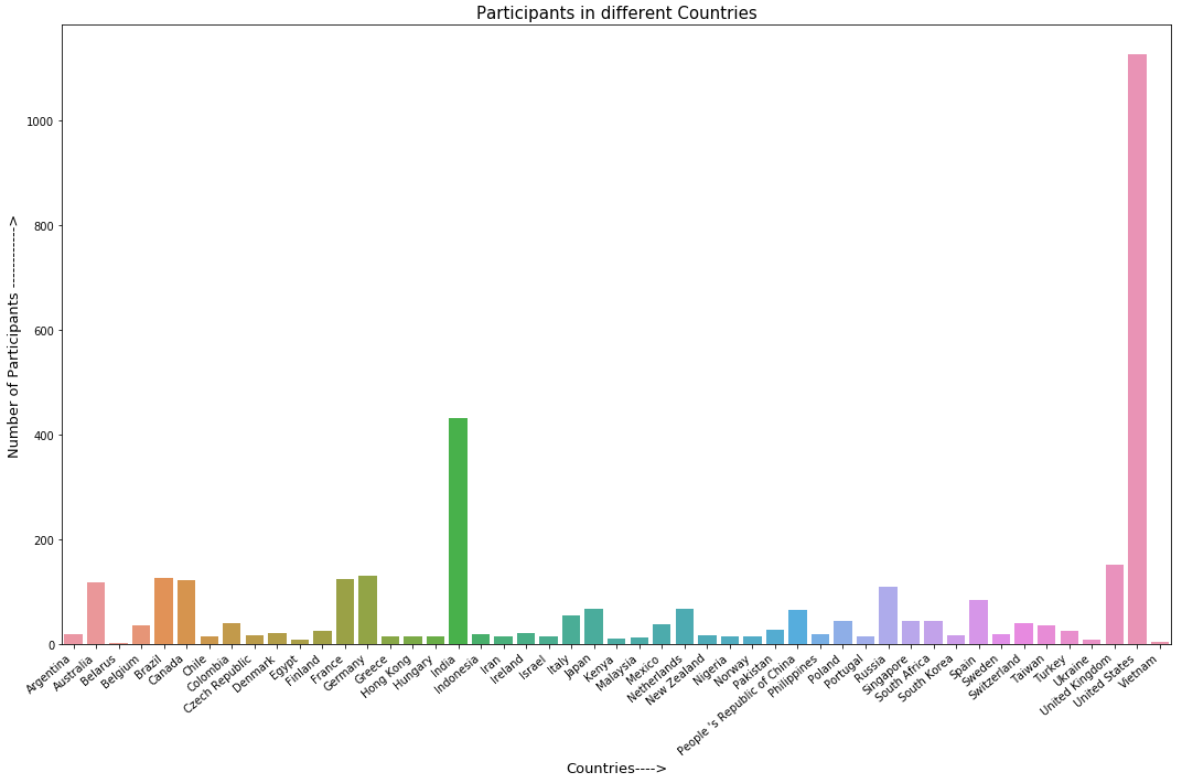
We know that people practise data science all over the world. The United States is thought of as a 'hub' of commercial data science as well as research followed by the United Kingdom and Germany.

Because the field is evolving so quickly, it may be that these perceptions, formed in the late 2000s are now inaccurate. So let's find out where data scientists live.

10. Create a bar graph of the respondants according to which country they are from.
Find the percentage of respondants from the top 5 countries

In [9]:

```
# Your code
fun={'Country':{'Country_count':'count'}}
country=response.groupby('Country').agg(fun).reset_index()
country.columns=country.columns.droplevel(0)
country.rename(columns={'':'Country_name'},inplace =True )
total_respondants= sum(country.Country_count)
#print(country)
plt.figure(figsize=(15,10))
ax=sb.barplot(x=country.Country_name,y=country.Country_count)
ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right",fontsize=10)
plt.title('Participants in different Countries',fontsize=15)
plt.xlabel('Countries---->',fontsize=13)
plt.ylabel('Number of Participants ------------>',fontsize=13)
plt.tight_layout()
plt.show()
#sorting data frame by Country_count in decending order
country.sort_values(["Country_count"], axis=0, ascending=False, inplace=True)
country=country.reset_index()
#calculating percentage respondants from top 5 countries
print("The percentage respondants of top 5 countries are :")
for i in range(5):
    print(country.Country_name[i]," : ",round((country.Country_count[i]/total_respo
```




```
The percentage respondants of top 5 countries are :
United States  :  31.81 %
India  :  12.18 %
United Kingdom  :  4.27 %
Germany  :  3.67 %
Brazil  :  3.59 %
```

Answer The percentage of respondants from top 5 countries are : United States with 31.81 % , India with 12.18 % , United Kingdom with 4.27 % , Germany with 3.67 % and Brazil with 3.59 % .

11. What comments can you make about our previous comments on the United States, United Kingdom and Europe?

Are the majority of data scientists now likely to come from those countries?

Answer United States is still the country with most number of participants, however according to data calculated, India is the second country with most number of participants before United Kingdom and Germany followed by Brazil.

12. Now that we have another demographic variable, let's see if there is any relationship between country, age and gender. We are specifically interested in the United States, India, United Kingdom, Germany and of course Australia!
Write code to output the mean and median age for each gender for United States, India, United Kingdom, Germany and Australia.

Hint: You may need to create a copy or slice.

In [10]:

```
# Your Code
countries=['United States', 'India', 'United Kingdom', 'Germany', 'Australia']
new_df=response[response.Country.isin(countries)].reset_index()
fun={'Age':{'Average_age':'mean','Median_age':'median'}}
stat_df=new_df.groupby(['Country','GenderSelect']).agg(fun).reset_index()
stat_df.columns=stat_df.columns.droplevel(0)
stat_df.columns=['Country','Gender','Average_age','Median_age']
#stat_df.rename(columns[={'':'Country_name','':'Gender'}],inplace =True )
stat_df
```

Out[10]:

	Country	Gender	Average_age	Median_age
0	Australia	Female	35.000000	34
1	Australia	Male	37.158416	36
2	Germany	Female	31.428571	29
3	Germany	Male	36.629310	34
4	India	A different identity	22.000000	22
5	India	Female	29.061224	28
6	India	Male	29.553806	28
7	United Kingdom	A different identity	36.000000	36
8	United Kingdom	Female	33.636364	33
9	United Kingdom	Male	35.811024	33
10	United States	A different identity	38.727273	43
11	United States	Female	34.370892	31
12	United States	Male	36.906874	34

13. What Pattern do you notice about the relationship between age, gender for each of these countries?

Answer In all the five countrie, we see that the average age of female is less that of male except in India, where both Men and Women have almost equal Average age(28 years) and Different Identity gender has low average age of 22 years.In the United States, participants with different identity are having higher average age(39 years) than other working males and females in the country. In United Kingdom also different Identity gender has higher average age alonge with male(36 years) than female(33).

2. Education

So far we have seen that there may be some relationships between age, gender and the country that the respondents are from. Next, we should look at what their education is like.

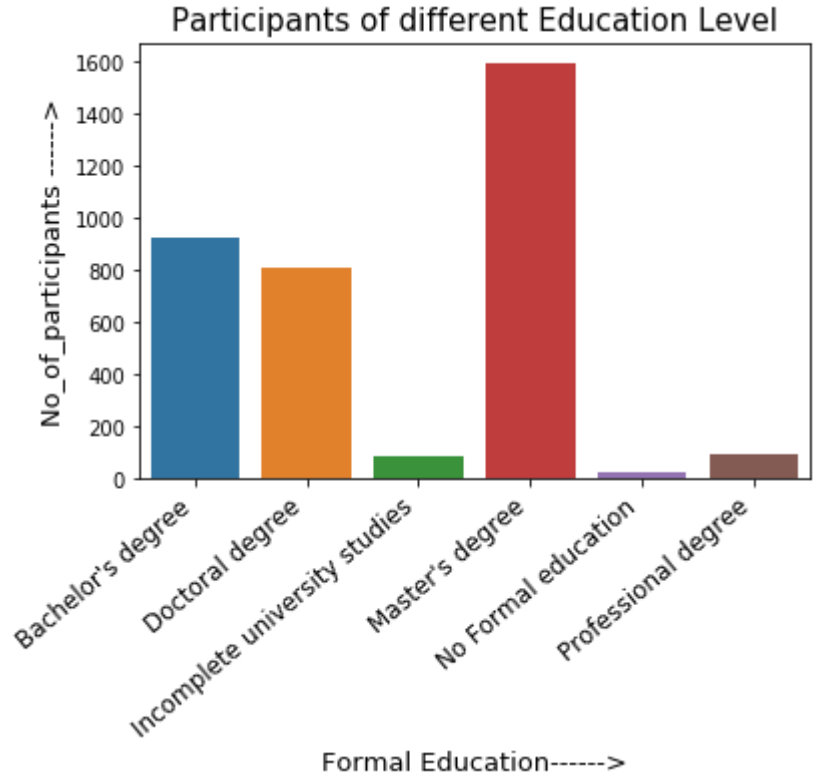
2.1 Formal education

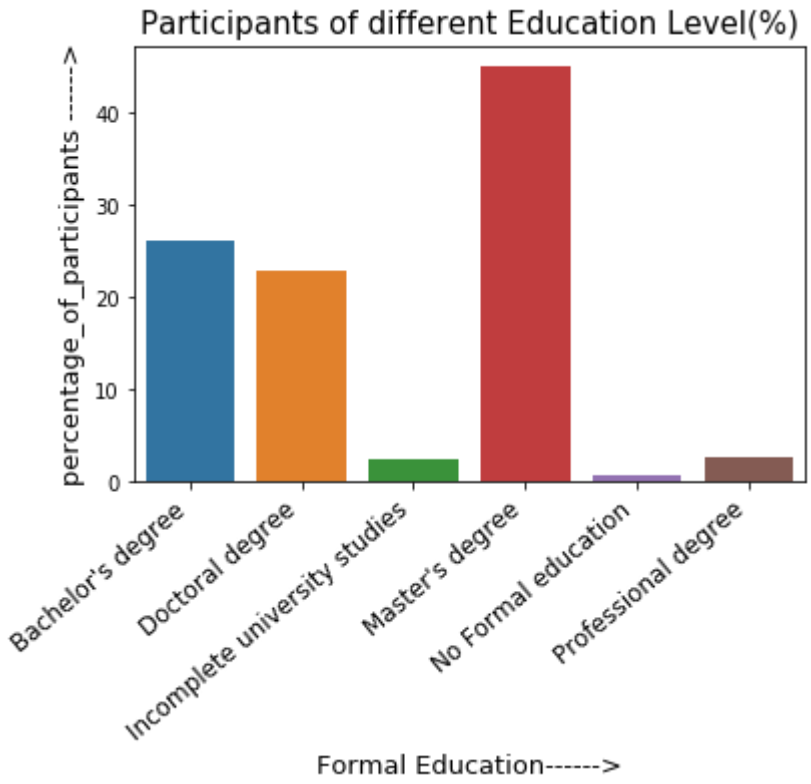
We saw in a recent activity that a significant number of job advertisements call for a masters degree or a PhD. Let's see if this is a reasonable ask based on the respondent's formal education.

14. Plot and display as text output the number and percentage of respondents with each type of formal education.

In [11]:

```
# Your code
# grouping data for counting partcipants with different education type
fun={'FormalEducation':{'No of participants':'count'}}
education_df=response.groupby('FormalEducation').agg(fun).reset_index()
education_df.columns=education_df.columns.droplevel(0)
education_df.columns=['FormalEducation','No_of_participants']

#plotting number of participants
ax=sb.barplot(x=education_df.FormalEducation,y=education_df.No_of_participants)
ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right",fontsize=12)
plt.title('Participants of different Education Level',fontsize=15)
plt.xlabel('Formal Education------>',fontsize=13)
plt.ylabel('No_of_participants ------>',fontsize=13)
plt.show()

#plotting percetage of participants
ax=sb.barplot(x=education_df.FormalEducation,y=(education_df.No_of_participants/tot
ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right",fontsize=12)
plt.title('Participants of different Education Level(%)',fontsize=15)
plt.xlabel('Formal Education------>',fontsize=13)
plt.ylabel('percentage_of_participants ------>',fontsize=13)
plt.show()

#text output

edu_len=len(education_df) # number of formal degree type
print("The number of Participants of different Education Level are :")
for i in range(edu_len):
    print(education_df.FormalEducation[i]," : ",education_df.No_of_participants[i])

print("The Percentage of Participants of different Education Level are :")
for i in range(edu_len):
    print(education_df.FormalEducation[i]," : ",round((education_df.No_of_participa
```






```
The number of Participants of different Education Level are :
Bachelor's degree  :  930
Doctoral degree  :  808
Incomplete university studies  :  87
Master's degree  :  1594
No Formal education  :  25
Professional degree  :  96
The Percentage of Participants of different Education Level are :
Bachelor's degree  :  26.27 %
Doctoral degree  :  22.82 %
Incomplete university studies  :  2.46 %
Master's degree  :  45.03 %
No Formal education  :  0.71 %
Professional degree  :  2.71 %
```

15. Based on what you have seen, do you think that a Master's or Doctoral degree is too unrealistic for job advertisers looking for someone with data science skills?

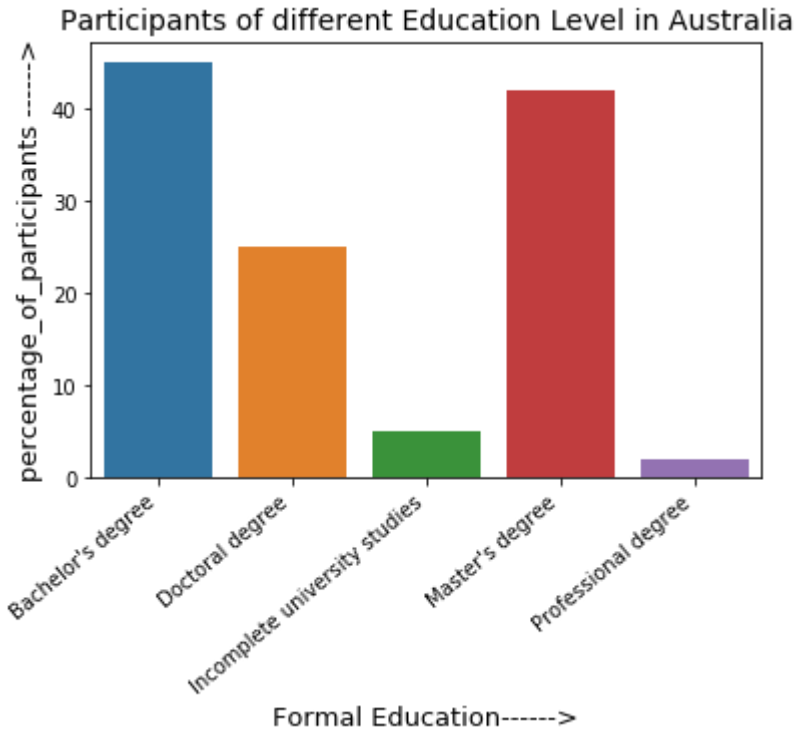
Give your reasons.

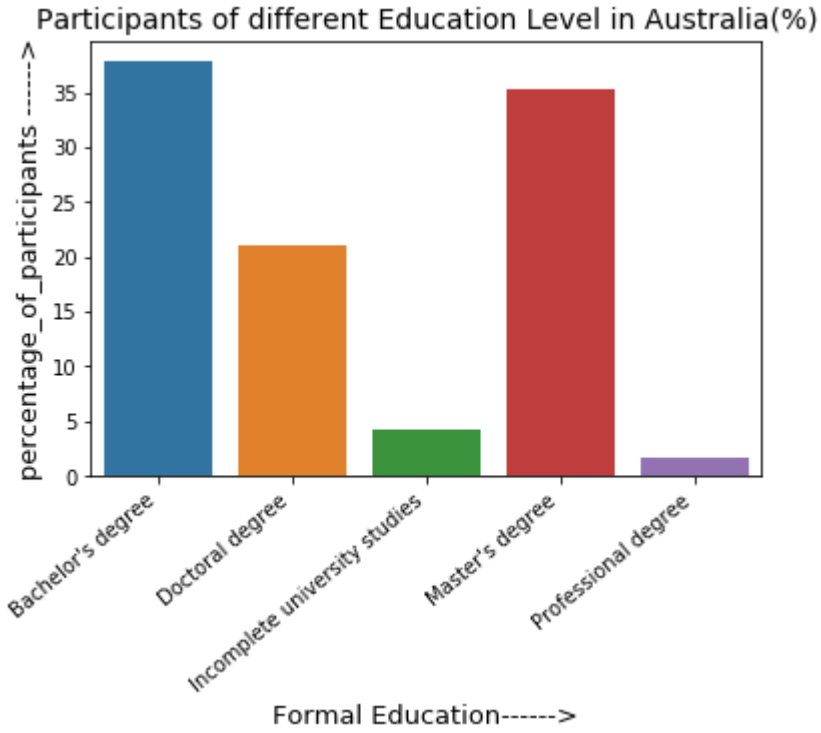
Answer No, I don't think Master's degree preference by Job advertiser is unrealistic, from the data we can see that more than 45% of the participants are holding Master's degree, this implies that Master's degree is higly considered and prefered by employers as this data of employed people.

16. Let's see if the trend is reflected in the Australian respondents.
Plot and display as text output the number and percentage of Australian respondents with each type of formal education.

In [12]:

```
# Your code
#filtering data for Australian participants
Aus_df=response[response.Country=='Australia'].reset_index()
total_res=len(Aus_df)
fun={'FormalEducation':{'No of participants':'count'}}
aus_edu_df=Aus_df.groupby('FormalEducation').agg(fun).reset_index()
aus_edu_df.columns=aus_edu_df.columns.droplevel(0)
aus_edu_df.columns=['FormalEducation','No_of_participants']

#plotting number of Australian participants Vs Education level
ax=sb.barplot(x=aus_edu_df.FormalEducation,y=aus_edu_df.No_of_participants)
ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right",fontsize=10)
plt.title('Participants of different Education Level in Australia',fontsize=14)
plt.xlabel('Formal Education------>',fontsize=13)
plt.ylabel('percentage_of_participants ------>',fontsize=13)
plt.show()

#plotting percentage of Australian participants Vs Education level
ax=sb.barplot(x=aus_edu_df.FormalEducation,y=(aus_edu_df.No_of_participants/total_r
ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right",fontsize=10)
plt.title('Participants of different Education Level in Australia(%)',fontsize=14)
plt.xlabel('Formal Education------>',fontsize=13)
plt.ylabel('percentage_of_participants ------>',fontsize=13)
plt.show()

#text output
aus_edu_len=len(aus_edu_df) # number of formal degree type
print("The number of Participants of different Education Level in Australia are :")
for i in range(aus_edu_len):
    print(aus_edu_df.FormalEducation[i]," : ",aus_edu_df.No_of_participants[i])

print("The Percentage of Participants of different Education Level in Australia are
for i in range(aus_edu_len):
    print(aus_edu_df.FormalEducation[i]," : ",round((aus_edu_df.No_of_participants[
```






```
The number of Participants of different Education Level in Australia a
re :
Bachelor's degree  :  45
Doctoral degree  :  25
Incomplete university studies  :  5
Master's degree  :  42
Professional degree  :  2
The Percentage of Participants of different Education Level in Austral
ia are :
Bachelor's degree  :  37.82 %
Doctoral degree  :  21.01 %
Incomplete university studies  :  4.2 %
Master's degree  :  35.29 %
Professional degree  :  1.68 %
```

17. Display as text output the mean and median age of each respondant according to each degree type.

In [13]:

```
# Your code
Mean_age_df=response.groupby('FormalEducation').mean()
Median_age_df=response.groupby('FormalEducation').median()
print("\nThe mean age of each type of degree is :\n",Mean_age_df.Age)
print("\nThe median age of each type of degree is :\n",Median_age_df.Age)
```

```
The mean age of each type of degree is :
 FormalEducation
Bachelor's degree                30.632258
Doctoral degree                  39.235149
Incomplete university studies    36.011494
Master's degree                  33.746550
No Formal education              41.680000
Professional degree              36.645833
Name: Age, dtype: float64

The median age of each type of degree is :
 FormalEducation
Bachelor's degree                28.0
Doctoral degree                  37.0
Incomplete university studies    35.0
Master's degree                  31.0
No Formal education              42.0
Professional degree              34.5
Name: Age, dtype: float64
```

3. Employment

After you complete your degree many of you will be seeking work. The graduate employment four months after graduation in Australia is 69.5%. At Monash, it is 70.1%. This is for all Australian degrees. Let's have a look at the state of the employment market for the respondents.

Let's have a look at the data.

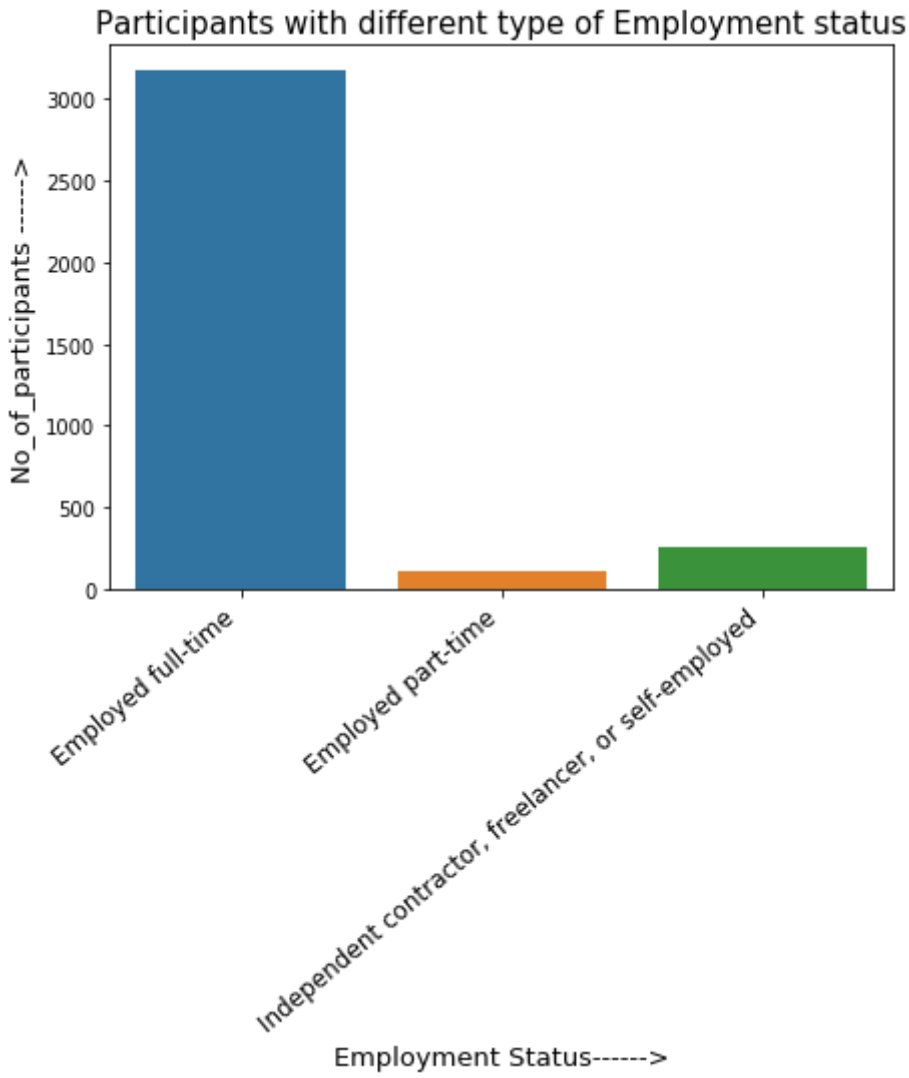
3.1 Employment status

The type of employment will affect the salary of a worker. Those employed part-time will likely earn less than those who work full time.

18. Plot the type of employment the respondents have on a bar chart.

In [14]:

```
# Your code
new_df=response.groupby('EmploymentStatus').count().reset_index()
plt.figure(figsize=(7,5))
ax=sb.barplot(x=new_df.EmploymentStatus,y=new_df.GenderSelect)
ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right",fontsize=12)
plt.title('Participants with different type of Employment status',fontsize=15)
plt.xlabel('Employment Status------>',fontsize=13)
plt.ylabel('No_of_participants ------>',fontsize=13)
plt.show()
```

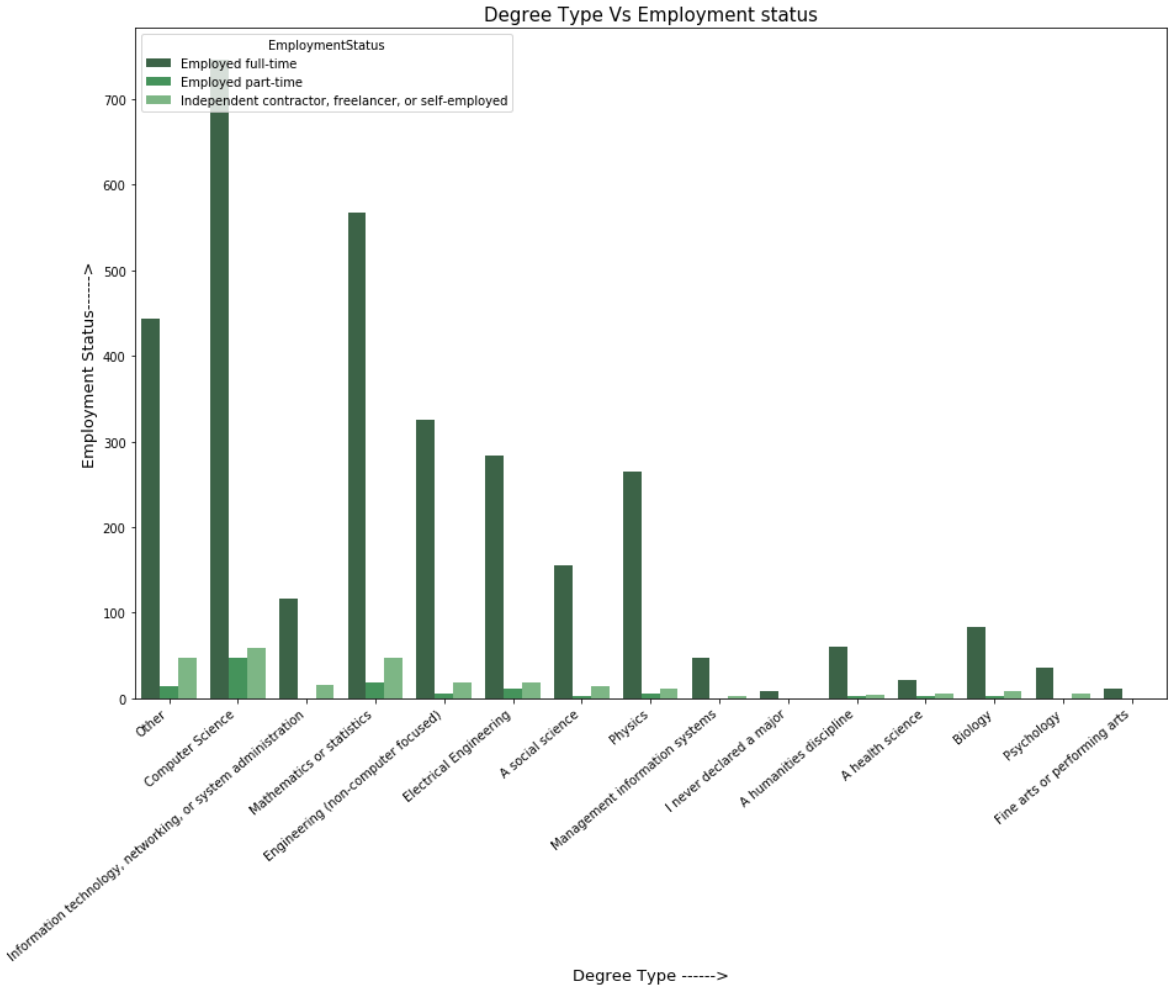


19. You may be wondering if your own degree and experince will help you gain full time employment after you graduate.

Plot the respondants employment types against their degrees.

In [15]:

```
# Your code
plt.figure(figsize=(15, 10))
ax=sb.countplot(x='MajorSelect', data=response, hue='EmploymentStatus',palette="Gre
ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right",fontsize=10)
plt.title('Degree Type Vs Employment status',fontsize=15)
plt.ylabel('Employment Status------>',fontsize=13)
plt.xlabel('Degree Type ------>',fontsize=13)
plt.show()
```



20. Looking at the graph, which degree is best to gain full-time employment?
What is odd about IT, networking or system administration??

Explain your answers.

Answer From the graph, we conclude that 'Computer Science' is the best degree to gain Full time employment with 854 participants out of 3540. 'IT,networking or system administration', a supposedly degree with computer science backgroud was expected to have more participants full time employment but the graph shows that there are only 133 participants with full time employment in this degree.

21. Overall, we know that 92.71% of respondents are employed, and 89.55% are employed full time. This may not be the same for every country. Print out the percentages of all respondents who are employed full time in Australia, United Kingdom and the United States.

In [16]:

```
# Your code
country=['United States', 'United Kingdom','Australia']
emp_df=response[response.Country.isin(country)]
emp_df=emp_df[emp_df.EmploymentStatus=='Employed full-time']
emp_df=emp_df.groupby('Country').agg({'EmploymentStatus':{'emp_count':'count'}}).re
emp_df.columns=emp_df.columns.droplevel(0)
emp_df.columns=['Country','emp_count']
for i in range(3):
    print("The percentages of all respondents who are employed full time in ", emp_
```

```
The percentages of all respondents who are employed full time in  Aust
ralia  are:  2.85 %
The percentages of all respondents who are employed full time in  Unit
ed Kingdom  are:  3.87 %
The percentages of all respondents who are employed full time in  Unit
ed States  are:  29.1 %
```

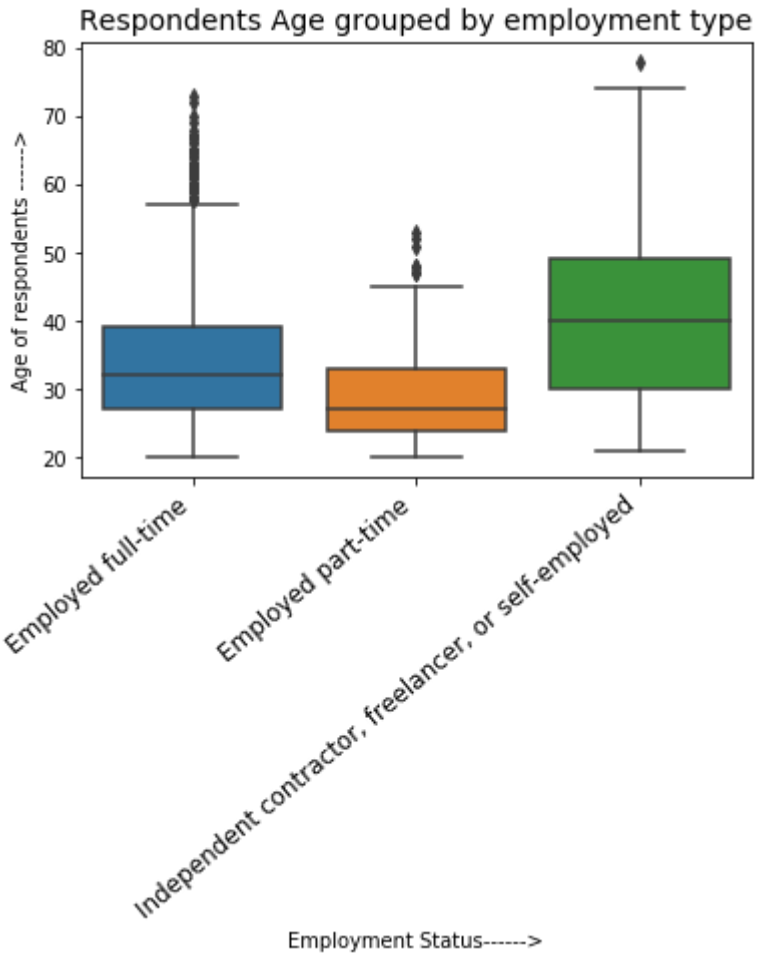
Remember earlier we saw that age seemed to have some interesting characteristics when plotted with other variables.

Let's find out the median age of employees by type of employment.

22. Plot a boxplot of the respondents age grouped by employment type.

In [17]:

```
# Your code
ax=sb.boxplot(x='EmploymentStatus',y='Age',data=response)
ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right",fontsize=12)
plt.title('Respondents Age grouped by employment type',fontsize=14)
plt.xlabel('Employment Status------>')
plt.ylabel('Age of respondents ------>')
plt.show()
```



Now this is interesting, full time employees seem to be a little older than part time employees. Independent contactors, freelancers and self-employed respondents are older still.

4. Salary

Data science is considered a very well paying role and was named 'best job of the year' for 2016.

We had a look around and saw that data scientists were paid between $110,823 at IBM and 149,963 at Apple, in Australian dollars.

On average it seems that $116,840 is what an Australian Data scientist can expect to earn. Do you think this is reasonable? Is this any different to the rest of the world?

4.1 Salary overview

Since all of the respondents did not come from one country, we can assume that they gave their salaries in their countries currency. We have filtered the data for you and provided exchange rates in a file called *conversionRates.csv* which should already be imported.

Let's have a look at the data.

23. Use the codes for each country to merge the files so that you can convert the salary data to Australian Dollars (AUD). Print out the maximum and median salary in AUD. Hint: think about what data type you have.

In [18]:

```
# Your code
conversion=pd.read_csv('conversionRates.csv')
exchange_df=pd.merge(response, conversion, how='inner', left_on=['CompensationCurre
len_of_df=len(exchange_df)
exchange_df['AUD_Salary'] =  (exchange_df['CompensationAmount'] * exchange_df['exch
print("The maximum Salary in AUD is",exchange_df['AUD_Salary'].max())
print("The median of Salary in AUD is",round(exchange_df['AUD_Salary'].median(),2))
```

```
The maximum Salary in AUD is 790290.0
The median of Salary in AUD is 76998.42
```

24. Do those figures reflect the values at the beginning of this section? Why do you think so?

Answer Yes, having maximum salary of AUD 790290 with median salary of AUD 76998 is the reason why in australia Data scientist earn upto AUD 116840 as according to data half of the participants earn more than AUD 76998 (median of salary).Australia is a developed nation, so it is expected to have salaries which is above average salary or median salary of the whole world.

4.2 Salary by country

Since each country has different cost of living and pay indexes, we should see how they compare.

25. Plot a boxplot of the Australian respondents salary distribution. Print out the maximum and median salaries for Australian repsondnts.

In [19]:

```
# Your code
sb.boxplot(y=response[response.Country=='Australia'].CompensationAmount)
plt.title('Salary distribution of Australians',fontsize=14)
plt.xlabel('Australian Respondents',fontsize=12)
plt.ylabel('Salary in AUD',fontsize=12)
plt.show()
print("The maximum salary for Australian repsondnts is ",response[response.Country=
print("The median of salary for Australian repsondnts is ",response[response.Countr
```


```
The maximum salary for Australian repsondnts is  500000.0
The median of salary for Australian repsondnts is  120000.0
```

26. Do those figures for Australia reflect the values at the beginning of this section?

Answer As given in the starting of the section that AUD 116840 is what an Australian Data scientist can expect to earn, the data shows that the median salary of australia is $120000, which is consistant to the data.

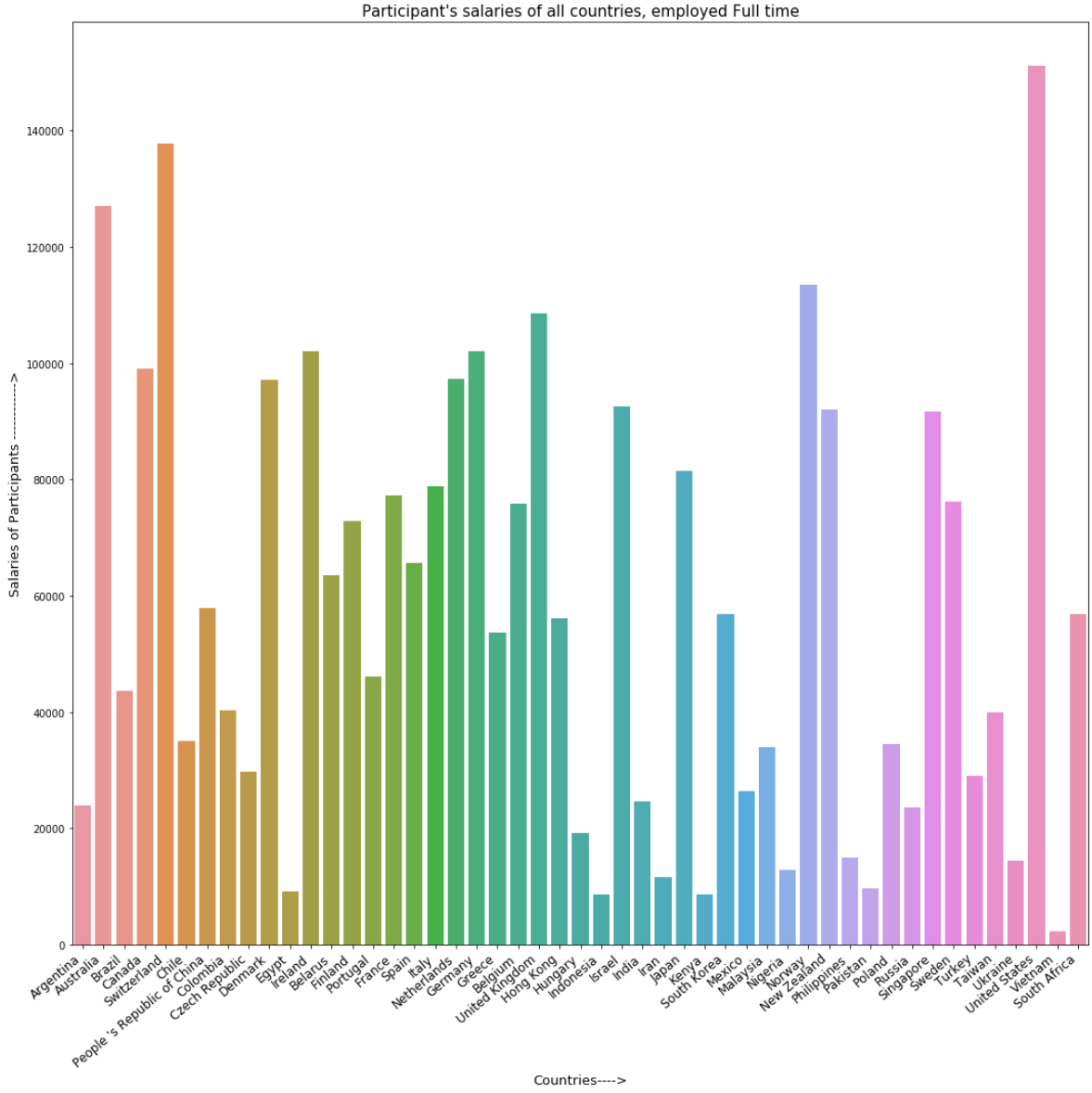
27. Australia's salaries look pretty good.
Plot the salaries of all countries on a bar chart.

Hint: Adjust for full-time employees only

In [20]:

```
# Your code
plt.figure(figsize=(15,15))
ax=sb.barplot(y=exchange_df[exchange_df.EmploymentStatus=='Employed full-time'].AUD
ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right",fontsize=12)
plt.title('Participant\'s salaries of all countries, employed Full time',fontsize=1
plt.xlabel('Countries---->',fontsize=13)
plt.ylabel('Salaries of Participants ------------>',fontsize=13)
plt.tight_layout()
plt.show()
```

```
/home/mranali/anaconda3/lib/python3.7/site-packages/scipy/stats/stats.
py:1713: FutureWarning: Using a non-tuple sequence for multidimensiona
l indexing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`.
In the future this will be interpreted as an array index, `arr[np.arra
y(seq)]`, which will result either in an error or a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```



28. What do you notice about the distributions? What do you think is the cause of this?

Answer The distribution of salary is uneven, countries like United States,United Kingdom, Australia,Switwerland and Germany have salary above the median salary on the other hand India and Brazil, also among the top employers, have salary way below than median.This could be due to the low currency rate, increasing population,Infrastructure and the demand for such job in these countries.

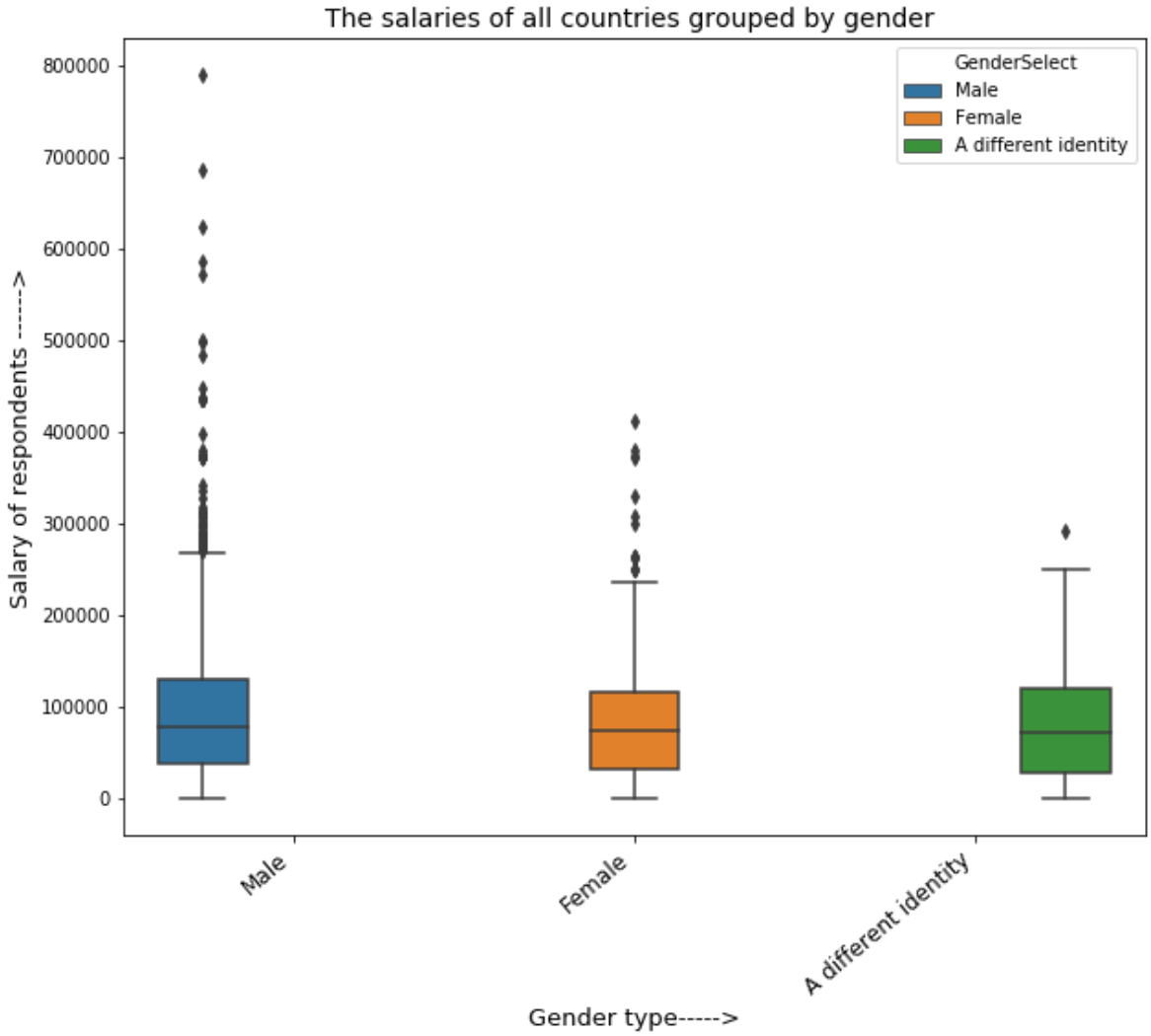
4.3 Salary and Gender

The gender pay gap in the tech industry is a big talking point. Let's see if the respondents are noticing the effect.

29. Plot the salaries of all countries grouped by gender on a boxplot.

In [21]:

```
# Your code
plt.figure(figsize=(10,8))
#ax=sb.boxplot(x='Country',y='AUD_Salary',data=exchange_df,hue='GenderSelect')
ax=sb.boxplot(x='GenderSelect',y='AUD_Salary',data=exchange_df,hue='GenderSelect')
ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right",fontsize=12)
plt.title('The salaries of all countries grouped by gender ',fontsize=14)
plt.xlabel('Gender type----->',fontsize=13)
plt.ylabel('Salary of respondents ------>',fontsize=13)
plt.show()
```



30. What do you notice about the distributions?

Answer The distribution is even for all the three gender types with males having slightly higher salary. Also, there are alot of outliers in boxplot for males implying that they are having higher salarie athan most of the males.

31. The salaries may be affected by the country the responant is from. In Australia the weekly difference in pay between men and women is 17.7% and in the United states it is 26%.
Print the median salaries of Australia, United States and India grouped by gender.

In [22]:

```
# Your code
country_list=['Australia','United States','India']
df=exchange_df[exchange_df.Country.isin(country_list)]
fun={'AUD_Salary':{'Median Salary':'median'}}
df=df.groupby(['Country','GenderSelect']).agg(fun).reset_index()
df.columns=df.columns.droplevel(0)
df.columns=['Country','Gender','Median Salary']
df
#print("The median salary of Australia grouped by gender is :",df.Country)
```

Out[22]:

	Country	Gender	Median Salary
0	Australia	Female	82000.000000
1	Australia	Male	130000.000000
2	India	A different identity	13628.148800
3	India	Female	12654.709600
4	India	Male	17327.217760
5	United States	A different identity	168264.137295
6	United States	Female	112176.091530
7	United States	Male	143336.116955

4.4 Salary and formal education

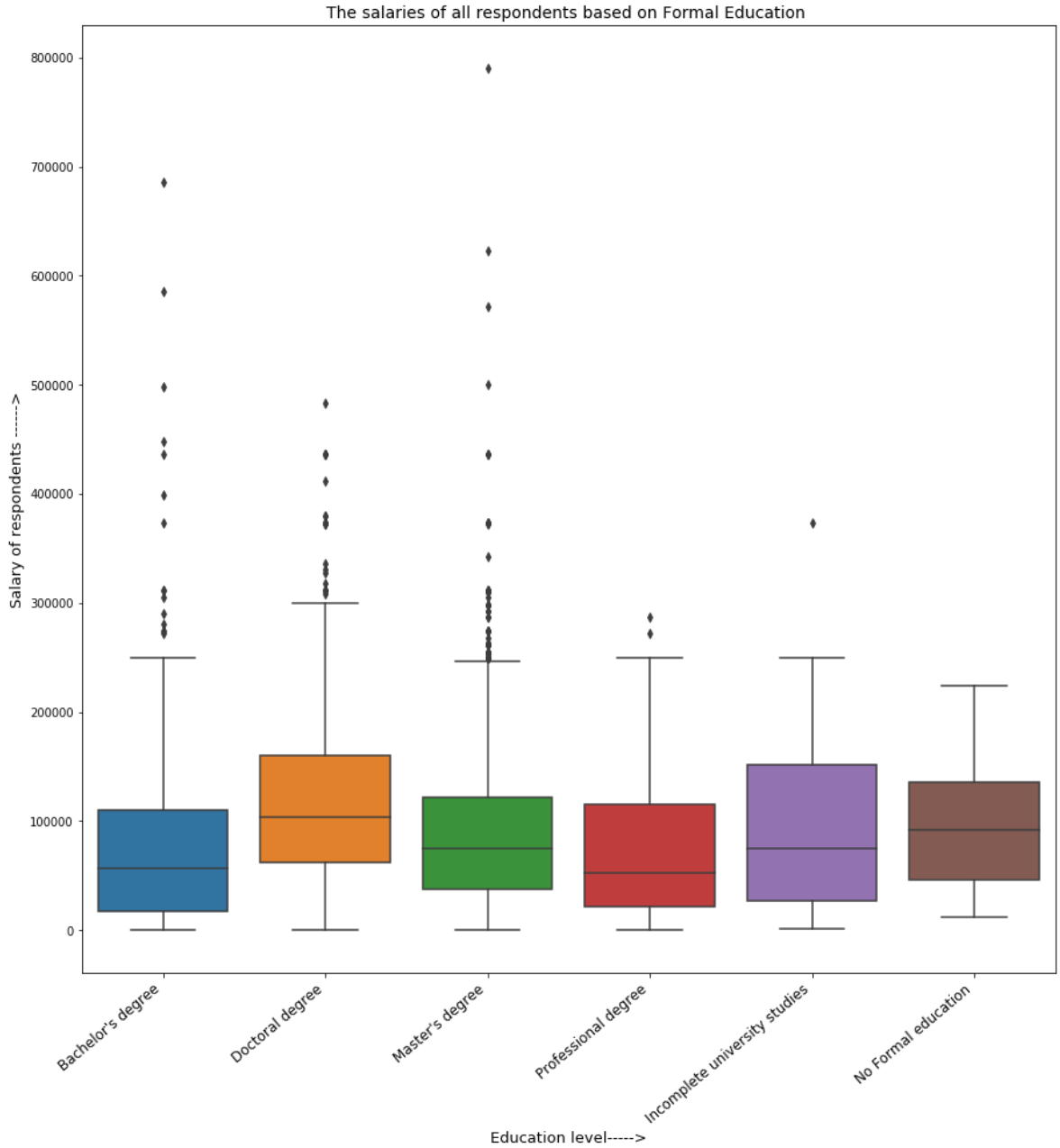
Is getting your master's really worth it ? Do PhDs get more money?

Let's see.

32. Plot the salary distribution of all respondants and group by formal education type on a boxplot.

In [23]:

```
# Your code
# Your code
plt.figure(figsize=(15, 15))
ax=sb.boxplot(x='FormalEducation',y='AUD_Salary',data=exchange_df)
ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right",fontsize=12)
plt.title('The salaries of all respondents based on Formal Education ',fontsize=14)
plt.xlabel('Education level----->',fontsize=13)
plt.ylabel('Salary of respondents ------>',fontsize=13)
plt.show()
```



33. Is it better to get your Masters or PhD?
Explain your answer.

Answer As per the graph, Doctoral degree or PhD degree has higher salary than Masters degree, based on

medians of nbothe grapth we can say that most of the Participants with PhD degree earn higher than participants with Master degree .So it is better to get PhD degree for better salary.However, considering outliers we can say that, though not majority but there are participants who earn more with Master degree and PhD.

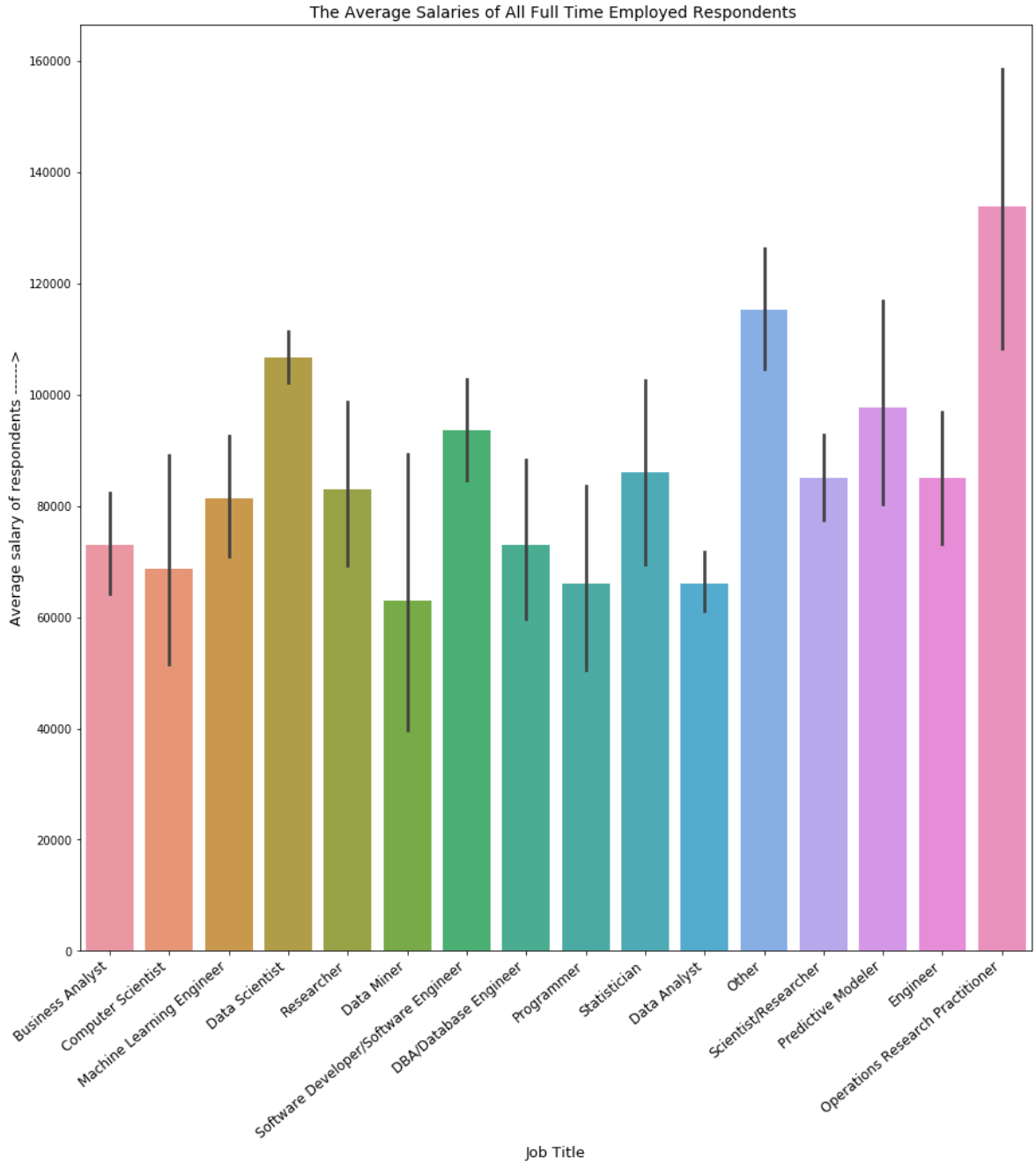
4.5 Salary and job

So are data scientists the highest paid in the industry? Or are there lesser known roles that are hiding from the spotlight?

34. Plot a bar chart of average salary (with error bars) of full time employees and group by job title.

In [24]:

```
# Your code
plt.figure(figsize=(15, 15))
fun={'AUD_Salary':{'Average Salary':'mean'}}
df=exchange_df[exchange_df.EmploymentStatus=='Employed full-time']
ax=sb.barplot(y=df.AUD_Salary,x=df.CurrentJobTitleSelect,ci=95,estimator=np.mean)
ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right",fontsize=12)
plt.title('The Average Salaries of All Full Time Employed Respondents ',fontsize=14
plt.xlabel('Job Title',fontsize=13)
plt.ylabel('Average salary of respondents ------>',fontsize=13)
plt.show()
```



35. Which job earns the most? Give a brief explaination of that job.

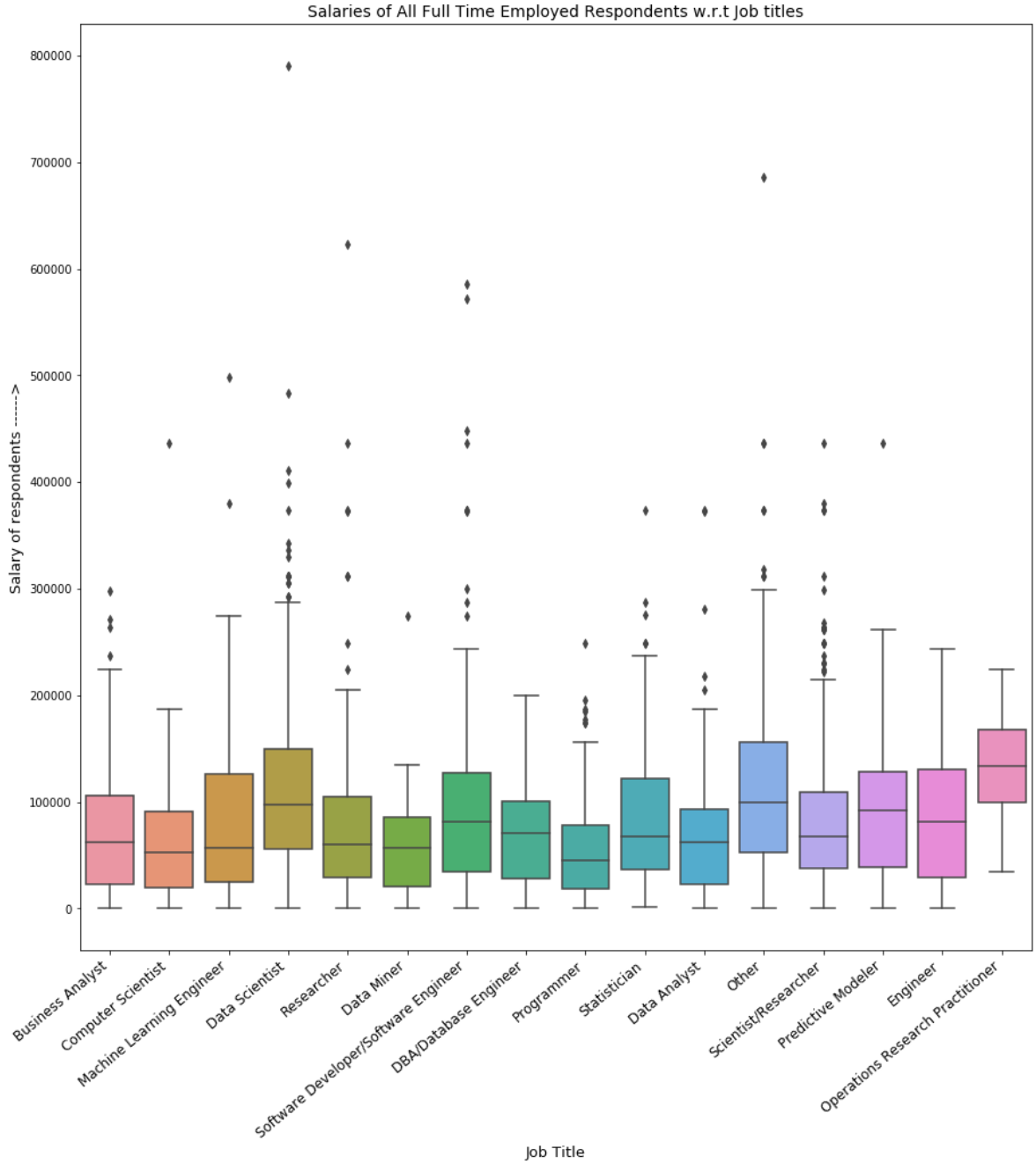
Answer An 'Operation Research Practitioner' gets highest paid. Operations Research is the study of how to

make decisions efficiently. Using mathematical Programming is one of the most powerful techniques used in Operations Research to the extent that sometimes both terms are used interchangeably. Operations Research practitioners solve real life problems that saves people money and time. These problems are very diverse and almost always seem unrelated. However, their essence is always the same, making decisions to achieve a goal in the most efficient manner.

36. So why are data scientists in the spotlight? Plot the salary distribution of full-time employees and group by job title as boxplots.

In [25]:

```
# Your code
plt.figure(figsize=(15, 15))
ax=sb.boxplot(y=df.AUD_Salary,x=df.CurrentJobTitleSelect)
ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right",fontsize=12)
plt.title('Salaries of All Full Time Employed Respondents w.r.t Job titles',fontsiz
plt.xlabel('Job Title',fontsize=13)
plt.ylabel('Salary of respondents ------>',fontsize=13)
plt.show()
```



37. Do the boxplots give some insight into why data scientists may receive so much attention? Explain your answer.

Answer According to the graph, participant who are Data Scientist have the higher pay among all other jobs except for participants in Other jobs(which can be business).The median of box plot for data Scientist is higher

than compare to other box plot medians.The reason could be that a Data Scientist have multiple skills which can related to any other job title and businesses today know the power of a good data analysis and prediction.

5. Predicting salary

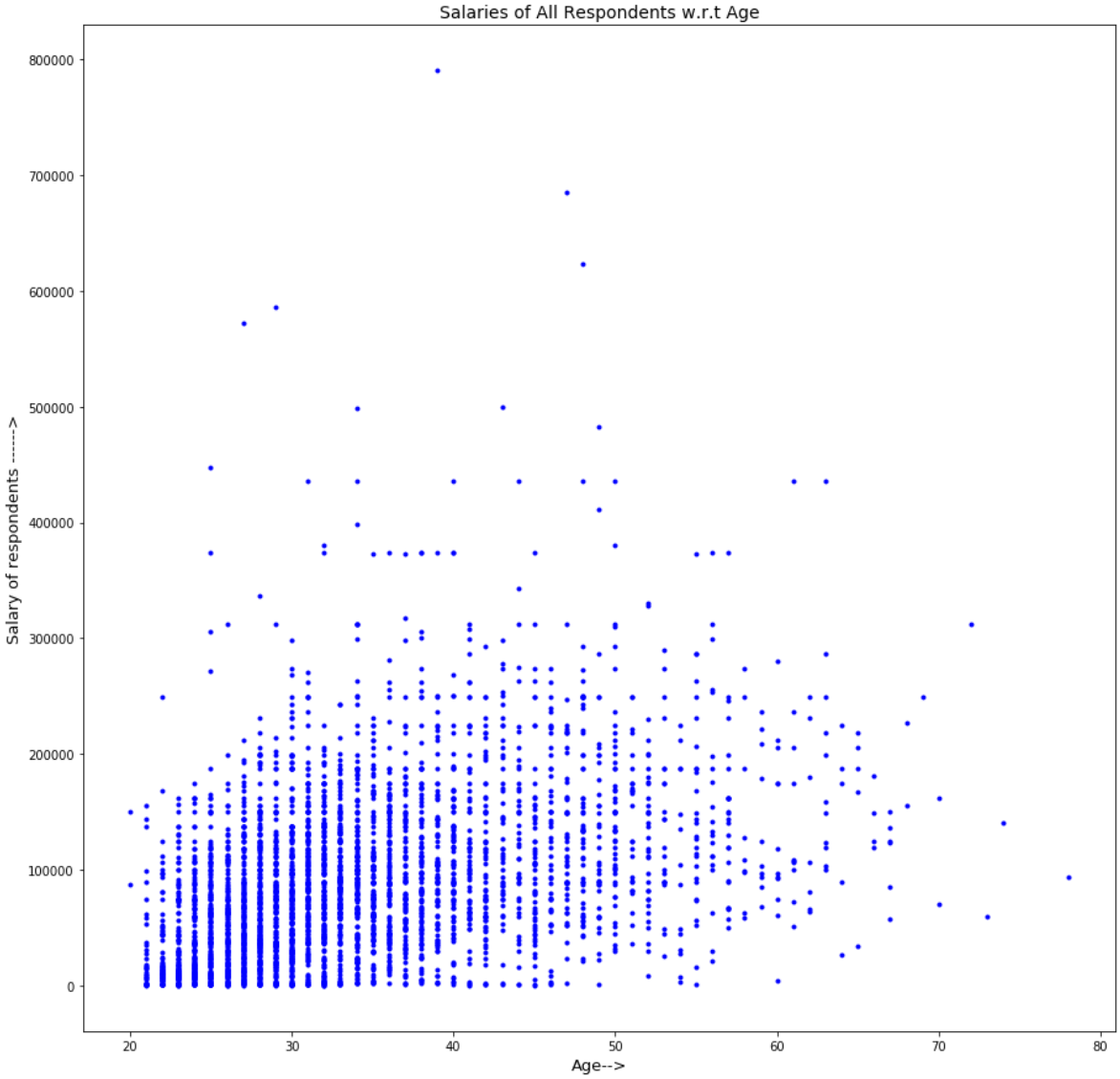
We have looked at many variables and seen that there are a lot of factors that could affect your salary.

Let's say we wanted to reduce it though? One method we could use is a linear regression. This is a very basic model that can give us some insights. Note though, there are more robust ways to predict salary based on categorical variables. But this exercise will give you a taste of predictive modelling.

38. Plot the salary distribution and age of respondants on a scatterplot.

In [26]:

```
# Your code
plt.figure(figsize=(15, 15))
plt.plot(exchange_df.Age,exchange_df.AUD_Salary,'.b')
plt.title('Salaries of All Respondents w.r.t Age',fontsize=14)
plt.xlabel('Age-->',fontsize=13)
plt.ylabel('Salary of respondents ------>',fontsize=13)
plt.show()
```

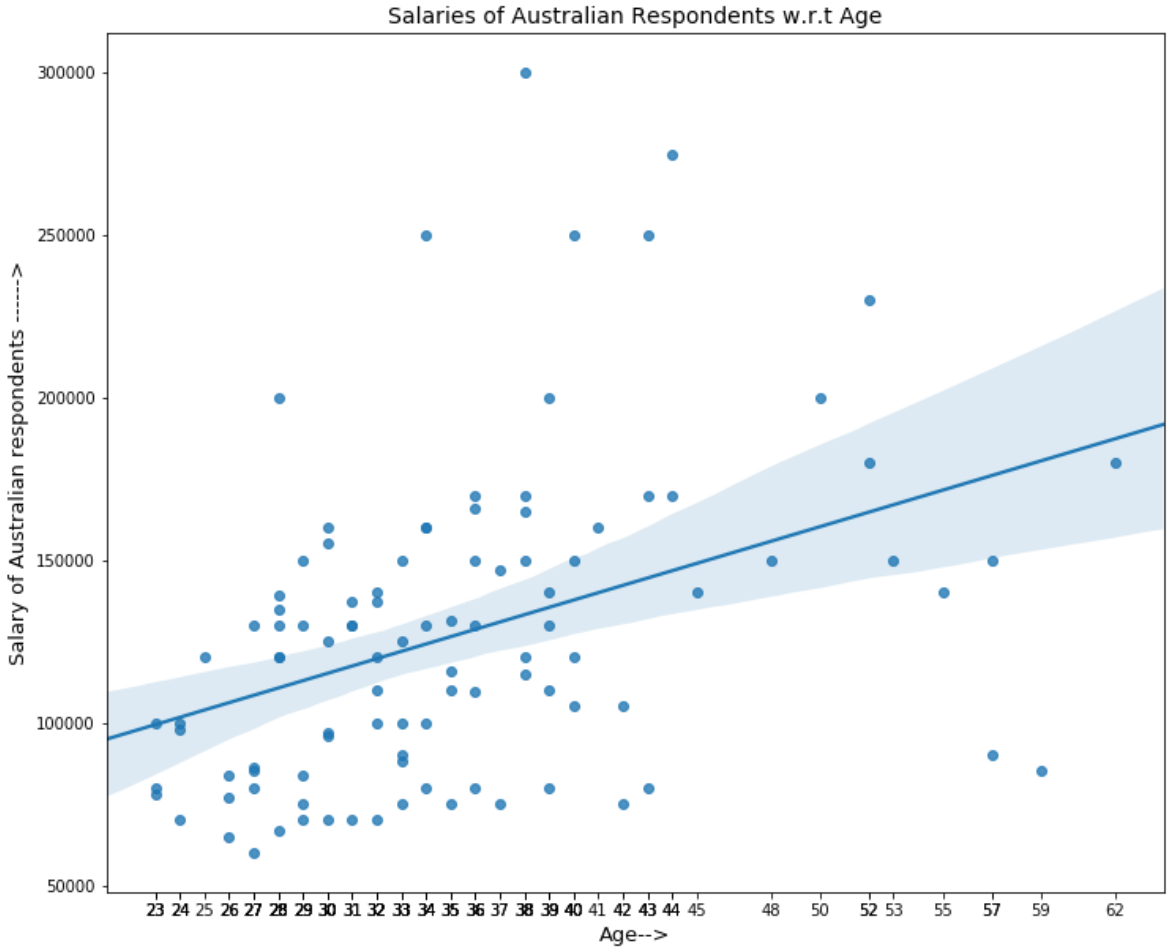


39. There may be a weak relationship. Let's refine this.

Create a linear regression between the salary and age of full-time Australian respondants. Plot the linear fit over the scatterplot.

In [27]:

```
#Your code
plt.figure(figsize=(12, 10))
df=exchange_df[exchange_df.Country=='Australia']
df1=df[df.EmploymentStatus=='Employed full-time']
ax=sb.regplot(x='Age',y='AUD_Salary',data=df1)
ax.set_xticks(df1.Age)
plt.title('Salaries of Australian Respondents w.r.t Age',fontsize=14)
plt.xlabel('Age-->',fontsize=13)
plt.ylabel('Salary of Australian respondents ------>',fontsize=13)
plt.show()
```



40. Do You think that this is a good way to predict salaries?
Explain your answer.

Answer No, it is not a good way, as we see that the linear line is not covering most of the scattered dots(implying that the realtion is not linear) and hence the we need another model which better to fits the relation between age and salary of the australian respondents.

Well done you have completed Part A. Don't forget Part B below.

For reassurance, the Graduate Careers Australia 2016 survey found the median salary for masters graduates in Computer Science and IT was $76,000.

Task B - Exploratory Analysis on Other Data

Find some publicly available data and repeat some of the analysis performed in Task A above. Good sources of data are government websites, such as: data.gov.au, data.gov, data.gov.in, data.gov.uk, ...

Please note that your report and analysis should contain consideration of the data you have found and its broader impact in terms of (1) the purpose of the data, (2) ethics and privacy issues, (3) environmental impact, (4) societal benefit, (5) health benefit, and (6) commercial benefit. Moreover, your analysis should at least involve (7) visualisation, (8) interpretation of your visualisation and (9) a prediction task.

To perform Task B, you can continue by extending this jupyter notebook file by adding more cells.

Analysis:

Data: Impact of various factors such as ethinicity, parent educational level,gender and test preparartion courses on Student Performance in Exam.

Data file source: https://www.kaggle.com/spscientist/students-performance-in-exams (https://www.kaggle.com/spscientist/students-performance-in-exams) The file 'StudentsPerformance.csv' has been taken from Kaggle databse, it has eight attributes, namely:

1.gender(Gender of students) 2.'race/ethinicity' (ethinic group students belog to) 3.'parental level of education' (education level of parents) 4.lunch (the amound/type of lunch students have in school) 5.test preparation course (Test preparation course taken?) 6.math score (marks of subject maths out of 100) 7.reading score (marks of reading out of 100) 8.writing score (marks of writing out of 100)

Inspiration

To study the impact of various factors, such as ethnic group,lunch type, taken preparation course and parental level of education on students performance in exams and to predict what category of students are likely to perform better in future.

In [28]:

```
#loading file
stu_per=pd.read_csv('StudentsPerformance.csv')
#Checking the file structure
stu_per.head()
```

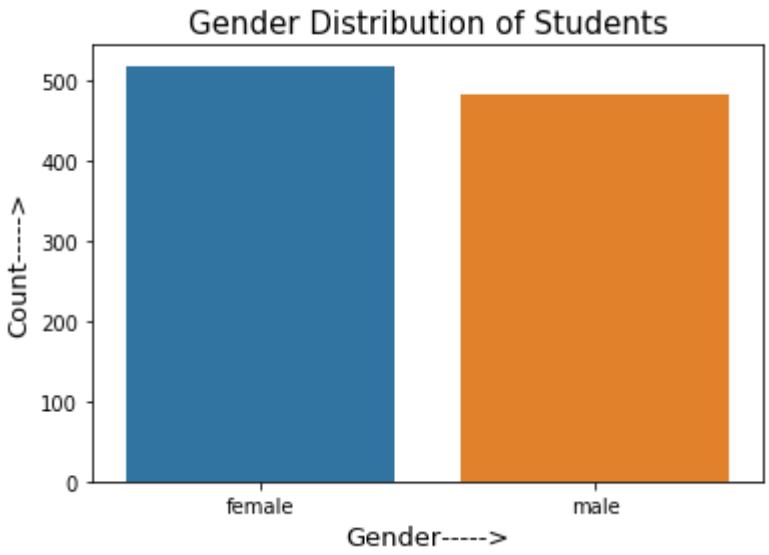
Out[28]:

	gender	ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75

1.Gender distribution in students

In [29]:

```
# cheching the gender distribution of all the students.
fun = {'gender' : {'gender_count':'count'}}
gen_cnt=stu_per.groupby('gender').agg(fun).reset_index()
gen_cnt.columns=gen_cnt.columns.droplevel(0)
gen_cnt.rename(columns={'':'Gender'},inplace =True )
sb.barplot(x=gen_cnt.Gender,y=gen_cnt.gender_count)
plt.title('Gender Distribution of Students',fontsize=15)
plt.xlabel('Gender----->',fontsize=13)
plt.ylabel('Count----->',fontsize=13)
plt.show()
```

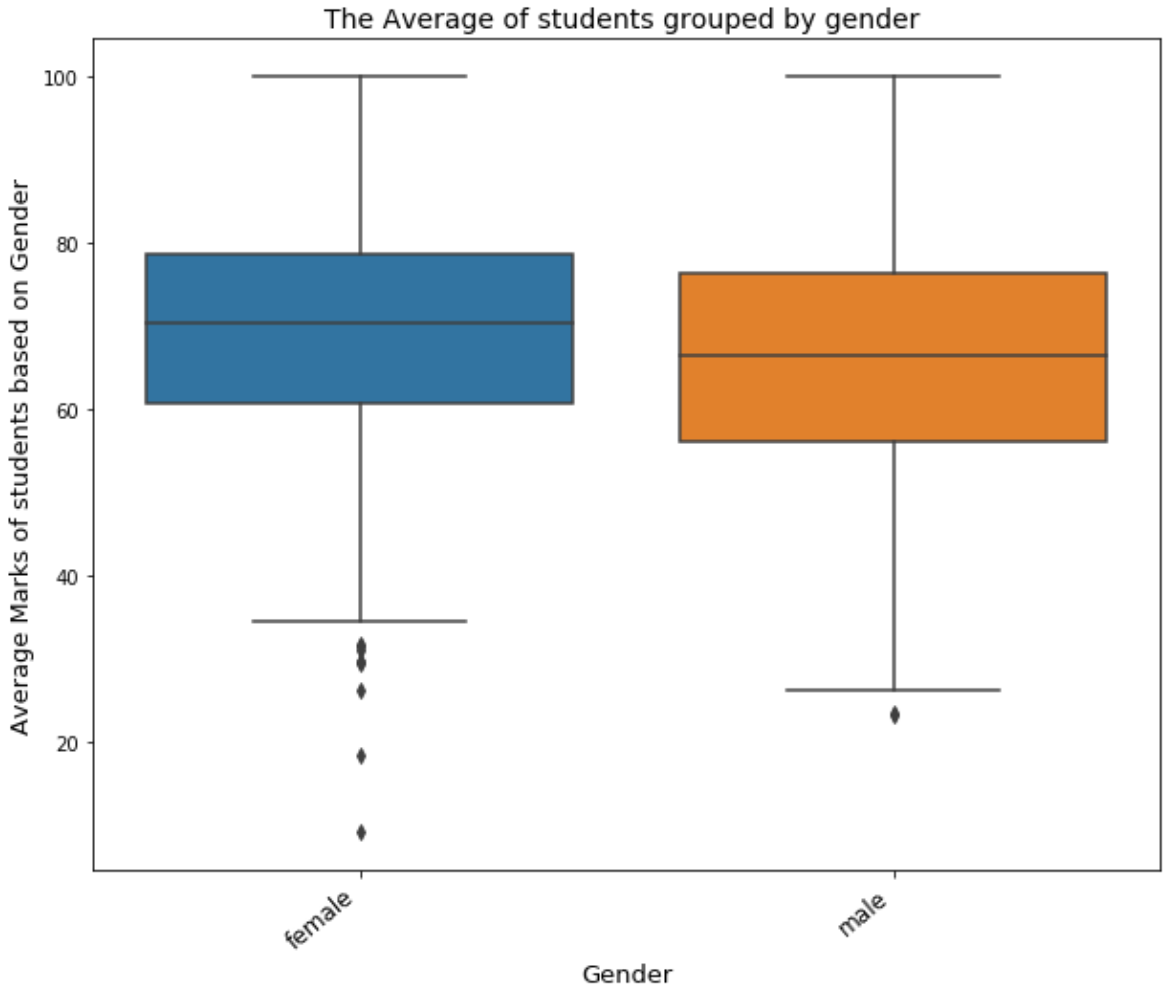


There are more female students than male students (518 female and 482 male students).

2. Plotting average marks of students grouped by gender.

In [30]:

```
#calculating average marks
plt.figure(figsize=(10,8))
stu_per['Average_marks']=((stu_per['math score']+stu_per['reading score']+stu_per['
ax=sb.boxplot(y=stu_per["Average_marks"],x=stu_per["gender"])
ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right",fontsize=12)
plt.title('The Average of students grouped by gender ',fontsize=14)
plt.xlabel('Gender',fontsize=13)
plt.ylabel('Average Marks of students based on Gender',fontsize=13)
plt.show()
```

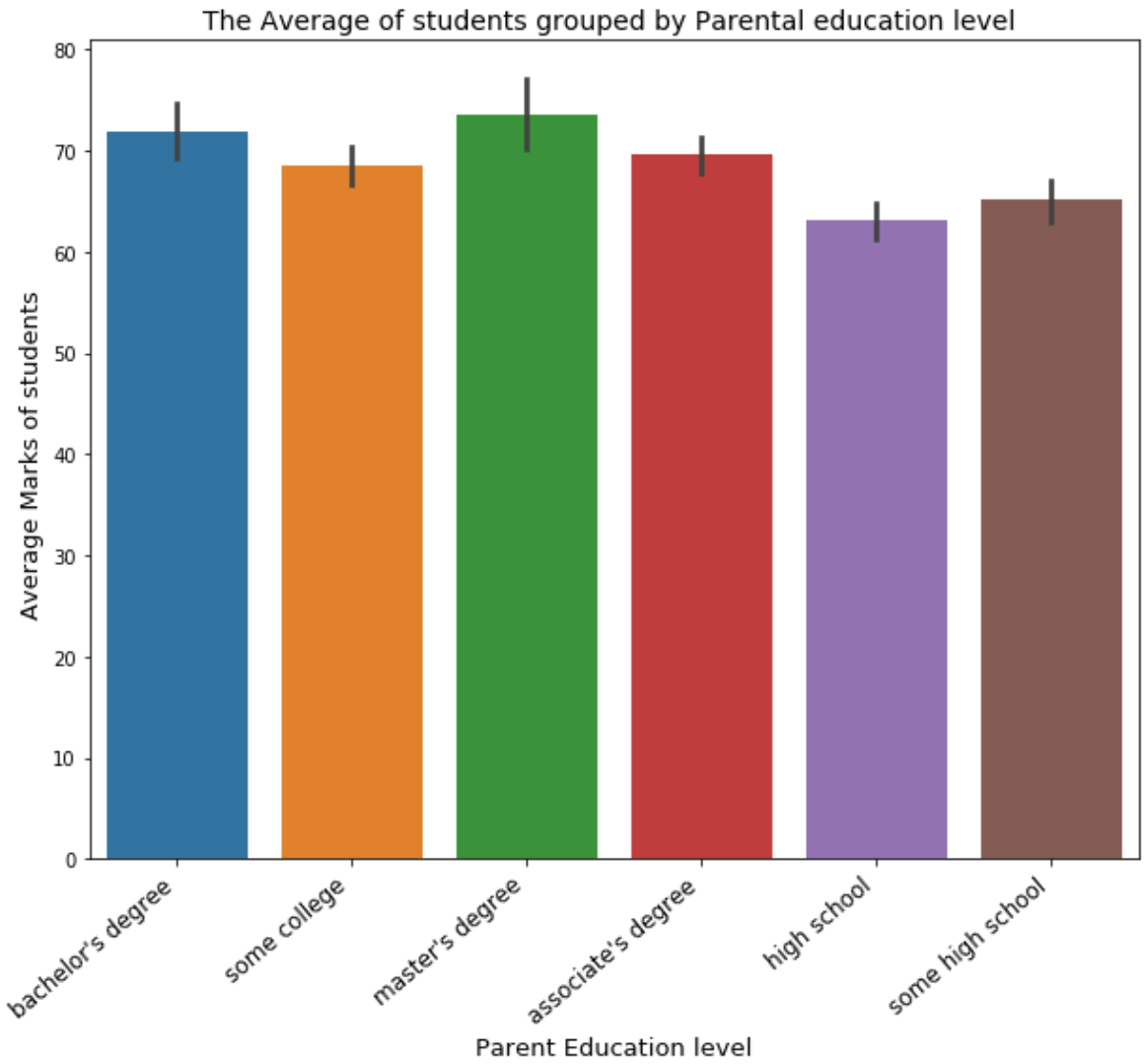


Box plot of female students has higher median and the 1st quantile is also higher as compared to male students implying that female students are having better average marks than male students.

3. Average marks of students grouped by Parental level of education.

In [31]:

```
plt.figure(figsize=(10,8))
ax=sb.barplot(y=stu_per["Average_marks"],x=stu_per["parental level of education"])
ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right",fontsize=12)
plt.title('The Average of students grouped by Parental education level ',fontsize=1
plt.xlabel('Parent Education level',fontsize=13)
plt.ylabel('Average Marks of students ',fontsize=13)
plt.show()
```



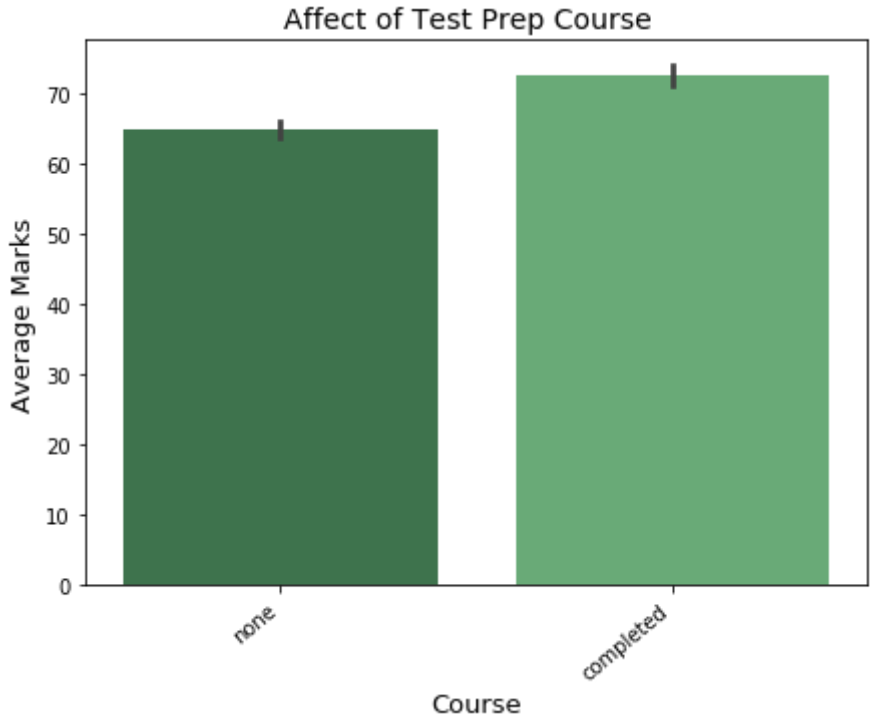
From the above graph we can say that, the average marks of student is directly proportional to parents education level as with increase in level of education of parents, the average marks of students increases, with highest in parents with Master's degree.

4. Is test preparation course worth?

Plotting average marks of students grouped by students who have taken the course and who have not.

In [32]:

```
plt.figure(figsize=(7,5))
ax=sb.barplot(y='Average_marks',x='test preparation course', data=stu_per, palette=
ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right",fontsize=10)
plt.title('Affect of Test Prep Course  ',fontsize=14)
plt.ylabel('Average Marks',fontsize=13)
plt.xlabel('Course',fontsize=13)
plt.show()
```

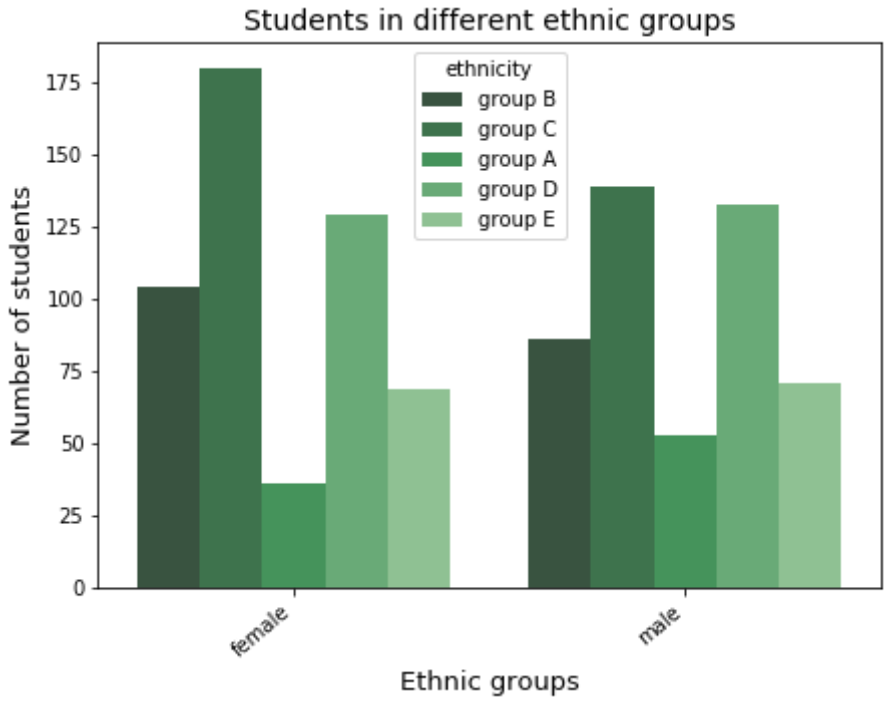


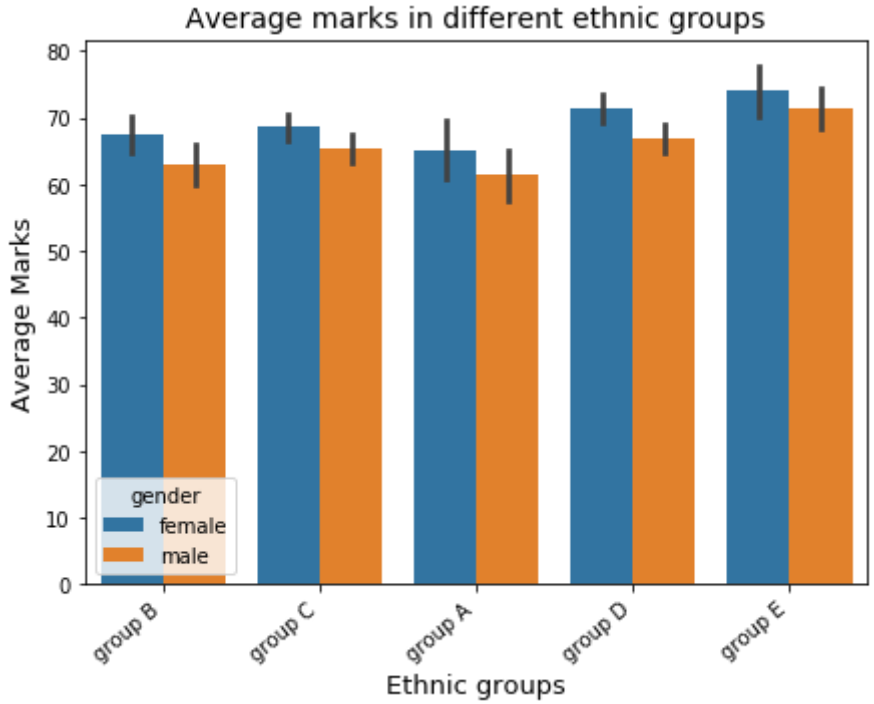
The above graph shows that taking test preparation course is beneficial for students as average marks of students who have taken the course is more than those who haven't.

5. Students distribution in different ethnicity

In [33]:

```
plt.figure(figsize=(7,5))
ax=sb.countplot(x='gender', data=stu_per, hue='ethnicity',palette="Greens_d")
ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right",fontsize=10)
plt.title('Students in different ethnic groups',fontsize=14)
plt.ylabel('Number of students',fontsize=13)
plt.xlabel('Ethnic groups',fontsize=13)
plt.show()

plt.figure(figsize=(7,5))
ax=sb.barplot(y='Average_marks', x='ethnicity',data=stu_per,hue='gender')
ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right",fontsize=10)
plt.title('Average marks in different ethnic groups',fontsize=14)
plt.ylabel('Average Marks',fontsize=13)
plt.xlabel('Ethnic groups',fontsize=13)
plt.show()
```





Based on above two graph we can say that group C has the highest number of students (with number of females more than number of males) and group A has the least number of students.From graph two we see that Students of ethnicity group E perform better than other and here also female students surpass male students.

6. Let's see if there is any relationship between ethnicity and subect score.

In [34]:

```
fun={'math score':{'avg_math_score':'mean'},'reading score':{'avg_reading_score':'m
new_df=stu_per.groupby('ethnicity').agg(fun).reset_index(0)
new_df.columns=new_df.columns.droplevel()
new_df.columns=['ethnicity','Avg_math_marks','Avg_reading_score',"Avg_writing_score
new_df
```

Out[34]:

	ethnicity	Avg_math_marks	Avg_reading_score	Avg_writing_score
0	group A	61.629213	64.674157	62.674157
1	group B	63.452632	67.352632	65.600000
2	group C	64.463950	69.103448	67.827586
3	group D	67.362595	70.030534	70.145038
4	group E	73.821429	73.028571	71.407143

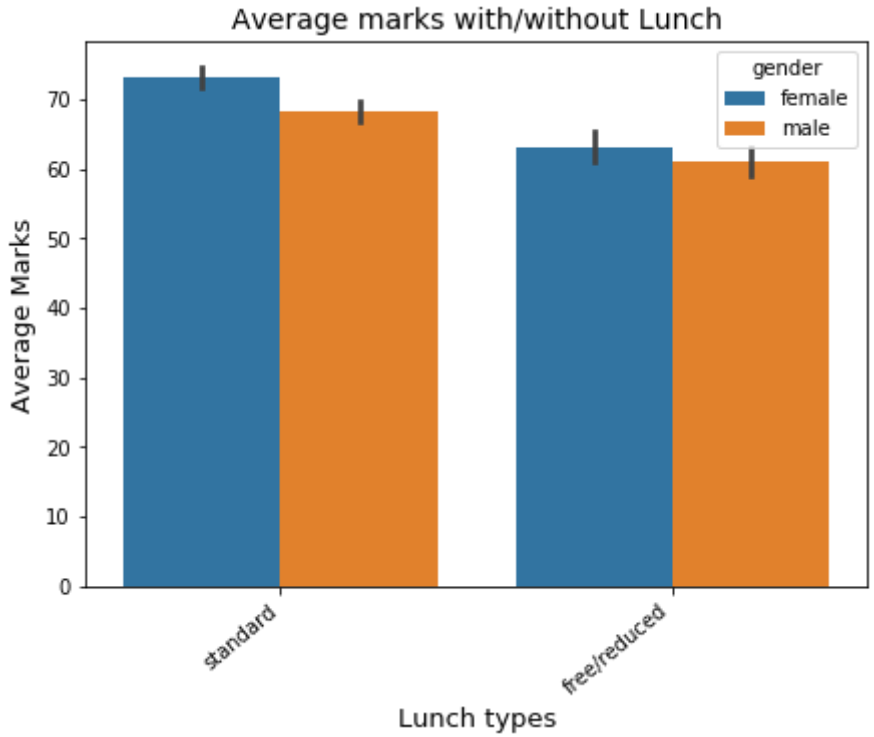
From above table, we conclude that individually also, group E has highest average marks in all the three subjects and group A has least marks.

7.Dietary importance

How having standard lunch and reduced lunch afftects the students performance.

In [35]:

```
plt.figure(figsize=(7,5))
ax=sb.barplot(y='Average_marks', x='lunch',data=stu_per,hue='gender')
ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right",fontsize=10)
plt.title('Average marks with/without Lunch',fontsize=14)
plt.ylabel('Average Marks',fontsize=13)
plt.xlabel('Lunch types',fontsize=13)
plt.show()
```

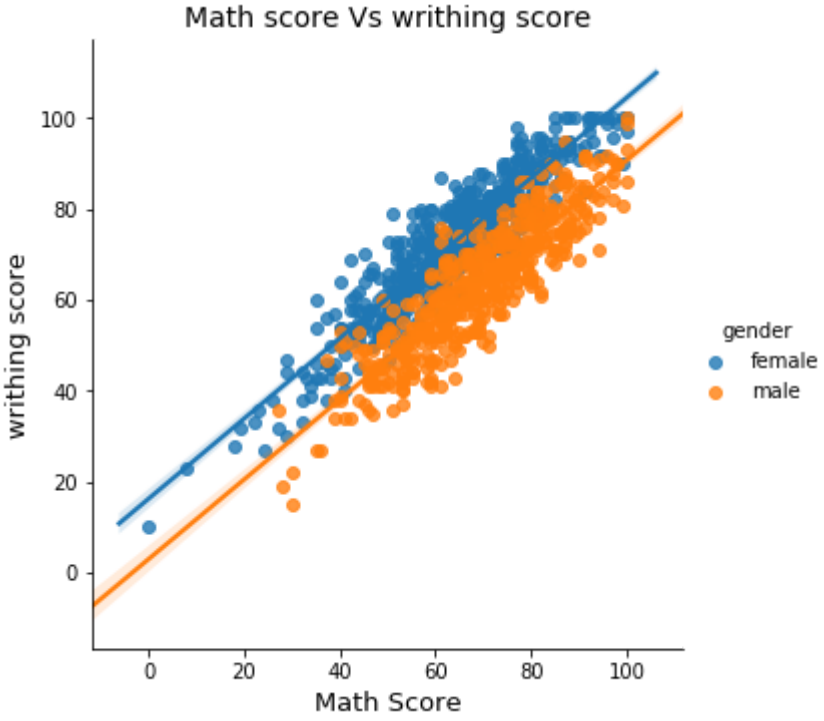


The figure shows that students without proper lunch get less marks and compared to ones with standard lunch/diet.

8.general prediction based on subject and gender of students

In [36]:

```
ax=sb.lmplot(x='math score',y='writing score',data=stu_per,hue='gender')
plt.title('Math score Vs writhing score',fontsize=14)
plt.xlabel('Math Score',fontsize=13)
plt.ylabel('writhing score',fontsize=13)
plt.show()
ax=sb.lmplot(x='math score',y='reading score',data=stu_per,hue='gender')
plt.title('Math score Vs Reading score',fontsize=14)
plt.xlabel('Math score',fontsize=13)
plt.ylabel('Reading Score',fontsize=13)
plt.show()
ax=sb.lmplot(x='reading score',y='writing score',data=stu_per,hue='gender')
plt.title('Reading score Vs Writhing score',fontsize=14)
plt.xlabel('Reading Score',fontsize=13)
plt.ylabel('writhing score',fontsize=13)
plt.show()
```



From above three scatter plot with linear model, we can say that Student who are good at reading are equally good at writing and that female students are better in all the three subjects than male students.This model can be used for predictions in future as the line almost fits the scattered plot.

In [ ]: