FIT5145 - Introduction to Data Science

Summer Semester B 2020

Assignment 1

This assesment aims to guide you in exploring a data set through the process of exploratory data analysis (EDA), primarily through visualisation of that data using various data science tools.

You will need to draw on what you have learnt and will continue to learn, in class. You are also encouraged to seek out alternative information from reputable sources. If you use or are 'inspired' by any source code from one of these sources, you must reference this.

Learning outcomes You will learn the following through completing this assessment:

1. Read in files and extract data from them into a data frame.
2. Wrangle and process data.
3. Use graphical and non-graphical tools to perform EDA.
4. Use basic tools for managing and processing big data.
5. Determine information
6. Communicate your findings in your report.

Submission details The Python code as a Jupyter notebook file (.ipyn). A PDF print of your Jupyter notebook containing the code, figures and answers to all the questions. Hint: Wrap your code using the Jupyter magics or pythonic standard.

Please note: Marks will be assigned based on their correctness and clarity of your answers and code. The PDF should be concise and not take up an excessive number of pages. You should not print the data frames in your PDF (comment out the code that prints those).

Zip file submissions attract a penalty of 10%. Submit two separate files requested above together. You will need to submit your PDF to Turnitin.

Task

In this course, you have learned about the definitions, skill sets, tools, applications and knowledge domains attributed to data science. However, these are extremely diverse and make data science challenging to define precisely. By completing the EDA, we hope you can get a clearer understanding of how a career in data science compares to others in the IT industry.

The Data

In late 2018, a survey was conducted for a large Australian collective of IT professionals. The survey, which received 7000 responses, aimed to gather information about IT professionals. The dataset was made public, and many insights have emerged since. We have taken 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.

The data set is called *assignment1_dataset.csv*, and contains respondents answers to survey questions. Each column contains the answers of one respondent to a specific question. Do not alter this dataset.

How to complete this assesment

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

Underneath the blue direction boxes, there are empty cells with the comment #Your code. Place your code in these. You should not need to but may 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. Write your answer under these.

Please note, your commenting and adherence to Python code standards will be marked. This notebook has been designed to give you a template for the layout of future notebooks you might create. 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.

Table of contents

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: Manali Choudhary

Student number: 30151198

Tutorial number. :P09

Tutor: Callum Ross Waugh

Environment: Python 3.7.4 and distribution (i.e. Anaconda 4.7.12 (64-bit))

Load your libraries and files

This assesment will be conducted using pandas. You will also be required to create visualisations. We recommend 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 on what each library does.

In [655]:

```
# Your code
 #import library for high level data manipulation and analysis
import pandas as pd
import matplotlib.pyplot as plt #import library for plotting
#import library for plotting- based on matplotlib
import seaborn as sns
#import library for linear regression calculations, line slopes etc.
from scipy.stats import linregress as lreg
#import library for maths operations and
#processing multidimentional array objects etc.
import numpy as np
#import library for data vizualising in wrodcloud
from wordcloud import WordCloud

sns.set(style="whitegrid")
sns.set(style="ticks")
dataset_a1 = pd.read_csv('assignment1_dataset.csv')
#dataset_a1
```

1. Demographic Analysis

Who are the survey participants?

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 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.

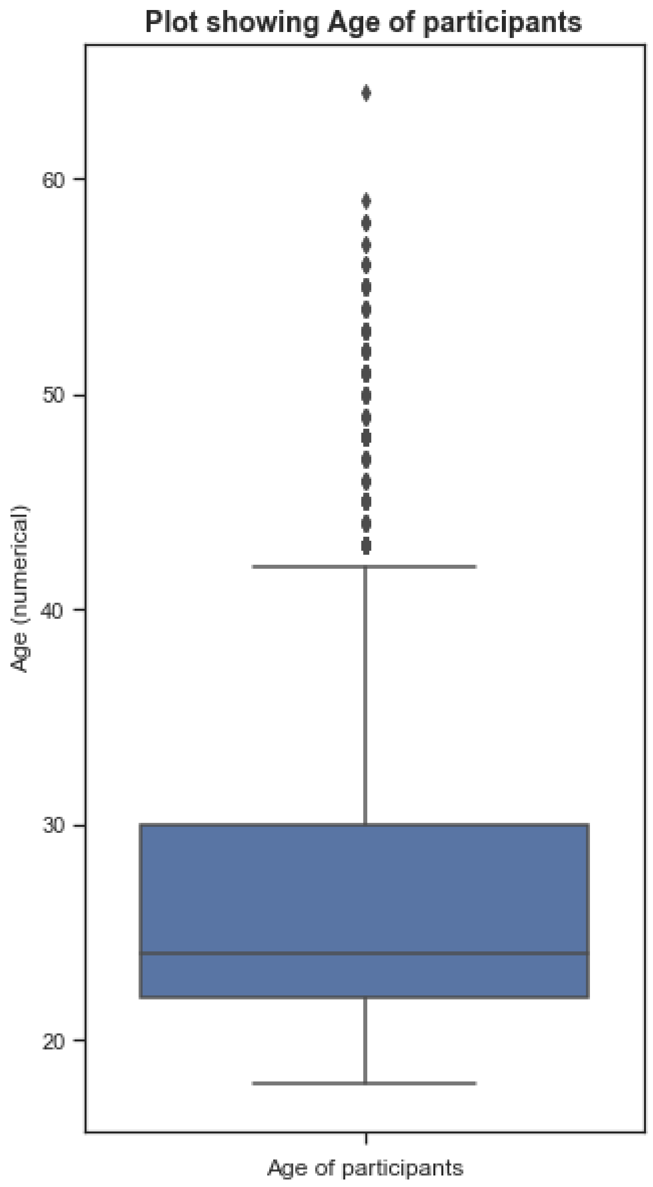
1. Create a box plot showing the age of all the participants.
Your plot must have labels for each axis, a title, numerical points for the age axis and also show the outliers.

In [603]:

```
# Your code
plt.figure(figsize=(5,10)) #set the fig size
ax = sns.boxplot(y="Age", data=dataset_a1) #plots the boxplot
#set the title in bold and particular fontsize and the axis labels
ax.set_title("Plot showing Age of participants",
             weight='bold').set_fontsize('14')
ax.set(xlabel = 'Age of participants',ylabel = 'Age (numerical)')
```

Out[603]:

```
[Text(0, 0.5, 'Age (numerical)'), Text(0.5, 0, 'Age of participants')]
```



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

In [680]:

```
# Your code
age = dataset_a1.Age
#calculate mean,min,max directly on age column
mean= np.mean(age)
min= np.min(age)
max= np.max(age)

#function to calculate Quartile 1
def q_1(x):
    a = np.percentile(x,25, interpolation='midpoint')
    return a.round().astype(int) #convert the answer to int

#calculate median by percentile function
median= np.percentile(age,50, interpolation='midpoint')
#function to calculate Quartile 3
def q_3(x):
    b = np.percentile(x,75, interpolation='midpoint')
    return b.round().astype(int)

#print the answer
print ("The mean and five descriptive statistics for age are: ", '\n'
        "Min:",min,"years", '\n'
       "Quartile 1:",q_1(age),"years",'\n'
       "Median:",median.round().astype(int),"years", '\n'
       "Quartile 3:",q_3(age),"years",'\n'
       "Mean:",int(round(mean)),"years",'\n'
       "Max:",max, "years")
```

```
The mean and five descriptive statistics for age are:
Min: 18 years
Quartile 1: 22 years
Median: 24 years
Quartile 3: 30 years
Mean: 27 years
Max: 64 years
```

Answer The minimum and maximum age of respondents recorded is 18 years and 64 years respectively.
The mean and median age of respondents recorded is 27 years and 24 years respectively.
The first quartile and third quartile age of respondents recorded is 22 years and 30 years respectively.

> 3.i. Looking at the boxplot, what general conclusion can you make about the age of the participants? You must explain your answer with reference to all five descriptive statistics. Simply listing will not suffice. You must discuss the conclusions drawn based on these descriptive statistics' relationship to each other. You must also make mention of the outliers if there are any.
>
> 3.ii. Would the mode be greater or lower than the mean? Why?

Answer i
According to the boxplot and five descriptives calculated,the line of median is near to the first quartile, lower than the third quartile,i.e. the longer part is above the median, hence we can say that, half the count of respondents are greater than or equal to 24 years of age i.e. youth, who are building their careers.By IQR- Q3-

Q1, the variability or interquartile range is 8 years. Hence,as the range of middle 50% of the data is not much but 8 years,the variability of data is not so great. The survey successfully targeted the respondents from minimum age of 18 years to maximum 64 years.i.e can say, from a beginners to the expert in their fields.As the max age according to the boxplot is around 42 years, but the descriptives show maximum age as 64 and the boxplot shows outliers specifically, outliers do exist beyond approx. 42 years of age.

Answer ii
According to the boxplot, we have already seen that the line of median is much lower than the third quartile,i.e. the longer part is above the median,hence it has positive/right skewness. Hence, we can determine that the mode would be lower than the mean and the median.

> 4. Regardless of the errors that the data show, we are interested in working-age IT professionals, aged between 20 and 65.
>
> Calculate how many respondents were under 20 or over 65?

In [380]:

```
# Your code
df = dataset_a1.Age

#filtering the dataframe using respective conditions
r_under20 = df < 20
r_over65 = df > 65

#print the count using the sum function on the derived series
print("Respondents under age '20':", sum(r_under20),'\n'
      "Respondents over age '65':", sum(r_over65))
```

```
Respondents under age '20': 90
Respondents over age '65': 0
```

Answer Number of respondents of the age under 20 is 90 and there are no respondents above the age 65 years.

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 that compares to our participants.

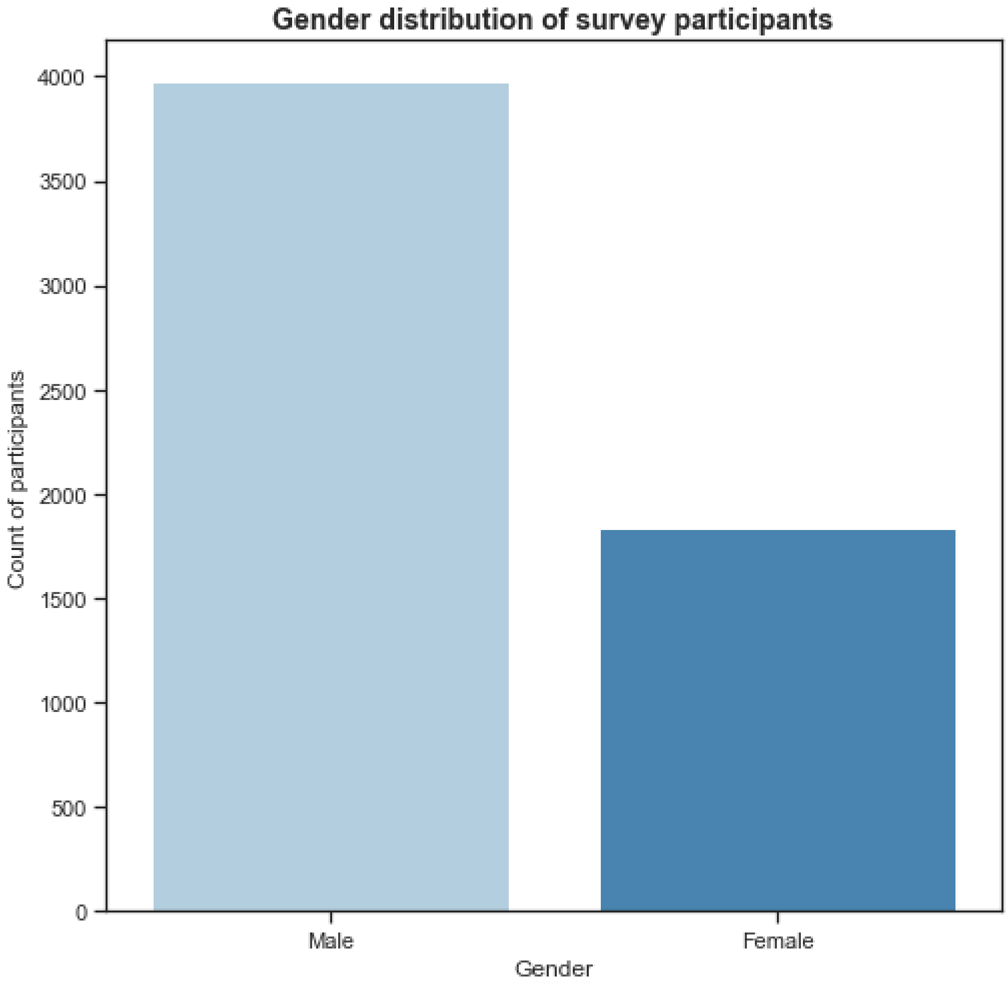
> 5. Plot the gender distribution of survey participants.

In [382]:

```
# Your code
plt.figure(figsize=(8,8))  #set the fig size

#plots the countplot with palette color blue
ax = sns.countplot(x = 'Gender',
                   data = dataset_a1, palette = 'Blues')

#set the title in bold and particular fontsize and the axis labels
ax.set_title('Gender distribution of survey participants',
             weight='bold').set_fontsize('14')
ax.set(ylabel = 'Count of participants')
```

Out[382]:

```
[Text(0, 0.5, 'Count of participants')]
```



6. Calculate what percentage of respondents were men and what percentage were women.

In [604]:

```python
# Your code
df = dataset_a1.Gender

#count the male and female respondents using gender column
c_male = df[df == 'Male'].count()
c_female = df[df == 'Female'].count()
total = df.count() #calculate the total count of the respondents

#function the calculate the percentage
def perc(a,b):
    percent = a/b * 100
    return round(percent,2)

#print the percentages of men & women using the created perc function
print('Percent of Men respondents:', perc(c_male,total),'%' '\n'
      'Percent of Women respondents:', perc(c_female,total),'%')
```

```
Percent of Men respondents: 68.35 %
Percent of Women respondents: 31.65 %
```

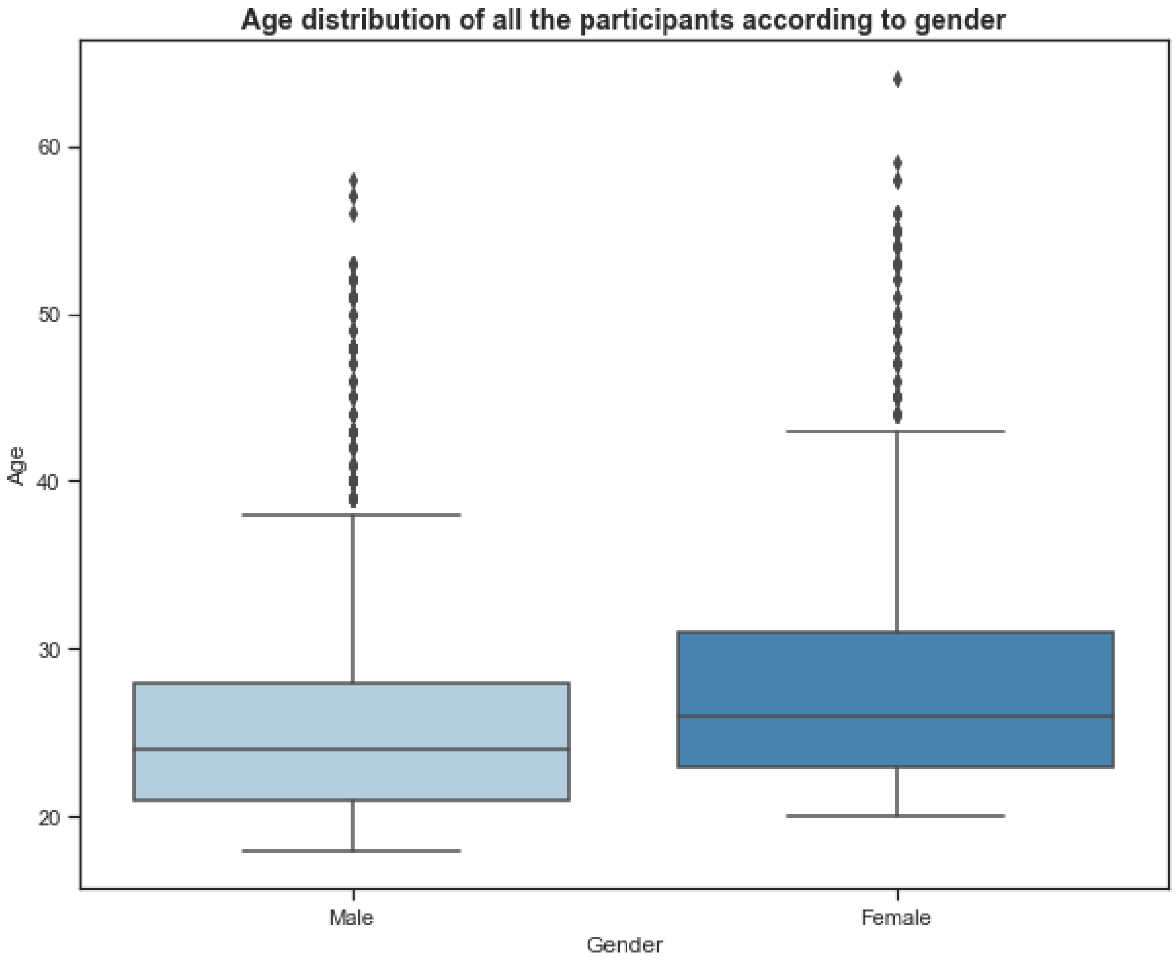
Answer

Percentage of male respondents is around 68% and that of female respondents is less than half of that of males i.e. around 32%

7. 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.

In [605]:

```
# Your code
plt.figure(figsize=(10,8)) #set the fig size

#plots the boxplot with palette color blue
ax = sns.boxplot(y="Age",x="Gender", data=dataset_a1,palette = 'Blues')

#set the title in bold and particular fontsize
ax.set_title(
    "Age distribution of all the participants according to gender",
             weight='bold').set_fontsize('14')
```



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

Hint: You need to determine the descriptive statistics.

In [653]:

```
# Your code
#calculate mean,min,max,median directly on age column
#used the previously created quartile function to calculate the quartile
fun = {'Age':{'Min':'min','Quartile 1':q_1,
              'Mean':'mean','Median':'median',
              'Quartile 3':q_3,'Max':'max'}}

#groupby Gender column and apply the agg function
groupbyGender = dataset_a1.groupby(
                            'Gender').agg(fun).round().astype(int)
groupbyGender = groupbyGender.reset_index() # reset its index
groupbyGender.columns = groupbyGender.columns.droplevel(0) # drop level 0 index
# rename the first column
groupbyGender.rename(columns = {'':'Gender'},inplace = True)
groupbyGender
```

Out[653]:

	Gender	Min	Quartile 1	Mean	Median	Quartile 3	Max
0	Female	20	23	28	26	31	64
1	Male	18	21	26	24	28	58

Answer

According to the boxplots, the distribution of male and female is positive/right skewed. The maximum age of both male and female show outliers exist as the boxplot representation do not match the descriptive analysis calculated. The minimum age of both genders show a difference of 2 years, hence we can do a abductive analysis that hiring age of female is higher than that of males. The maximum age of Female is greater to that of male, hence, female seem to work for longer time as compared to that of male.

1.3 Country

We know that people practice IT all over the world. The United States is thought of as a central 'hub' for commercial IT services as well as research followed by the United Kingdom and Germany.

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

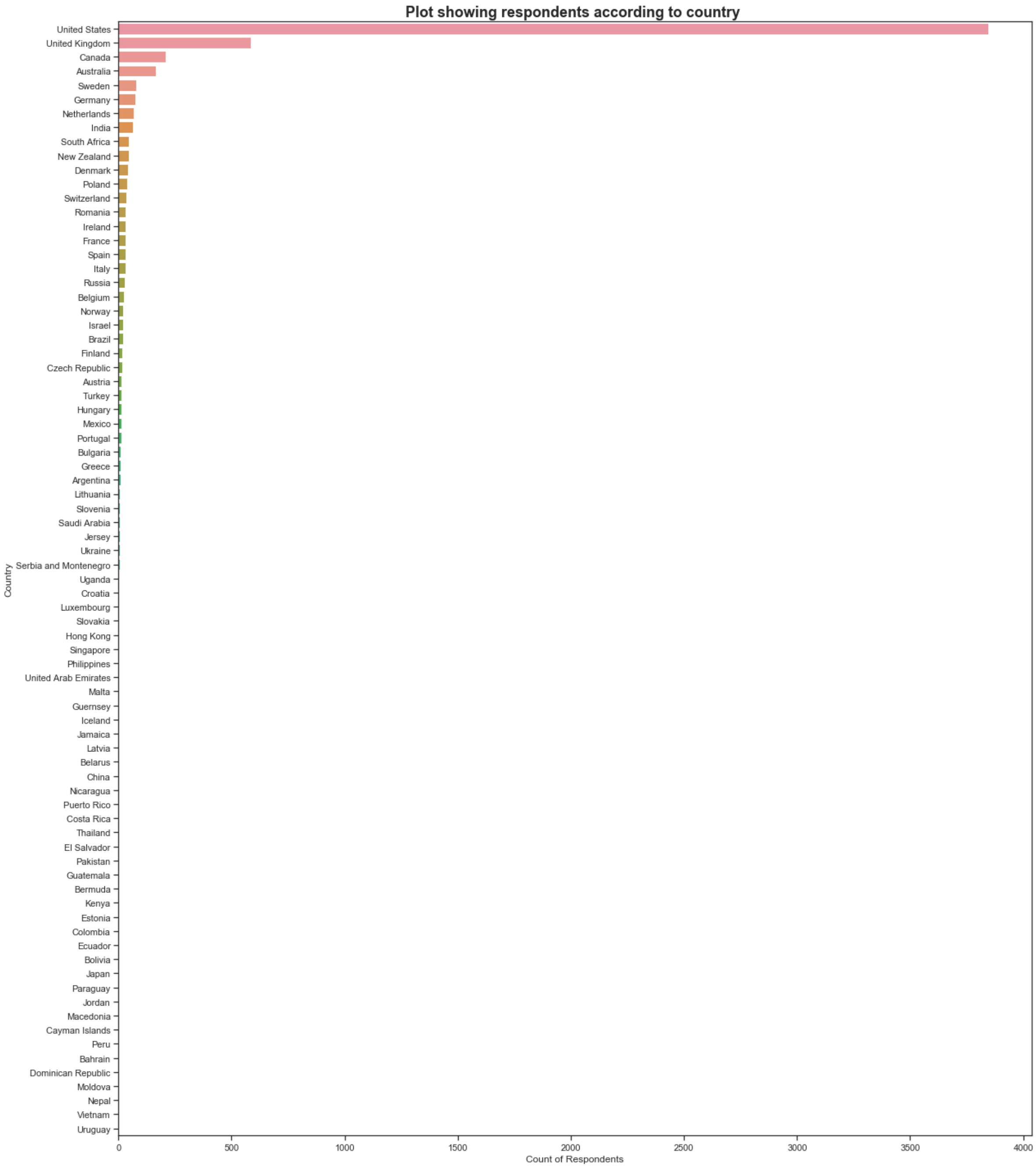
9. Create a bar graph of the respondents according to which country they are from.
Find the percentage of respondents from the top 5 countries.
Print your display rounding to two decimal places before writing out your answer.

In [609]:

```
# Your code
plt.figure(figsize=(20,25)) #set the fig size

#plots the countplot
ax = sns.countplot(y="Country", data = dataset_a1,
                   order = dataset_a1.Country.value_counts().index)

#set the title in bold and particular fontsize and the axis labels
ax.set_title("Plot showing respondents according to country ",
             weight='bold').set_fontsize('18')
ax.set(xlabel = 'Count of Respondents')
```

Out[609]:

```
[Text(0.5, 0, 'Count of Respondents')]
```



10. Find the percentage of respondents from the top 5 countries.
Print your display rounding to two decimal places before writing out your answer.

In [610]:

```
# Your code
country = dataset_a1.Country

#list of the values of the top 5 countries
country_list = list(country.value_counts().head(5))

#list of the names of the top 5 countries
country_names = list(country.value_counts().head(5).keys())
total = country.count() #total count of respondents of all countries

#iterate the country_list using map
#apply the perc function created previously on the values and store it as a list
percent_list = list(map(lambda x: perc(x, total),
                        country_list))

#create a new dataframe of the top 5 countries and
#their resp. percentages of respondents
df = pd.DataFrame({'Country' : country_names ,
                   'Percentage of respondents' : percent_list})
df
```

Out[610]:

	Country	Percentage of respondents
0	United States	66.15
1	United Kingdom	10.04
2	Canada	3.61
3	Australia	2.87
4	Sweden	1.32

@Muhammad Yasoob Ullah Khalid. (2017). Answer to question: Map, Filter and Reduce. Retrieved from: [https://book.pythontips.com/en/latest/map_filter.html (https://book.pythontips.com/en/latest/map_filter.html)](https://book.pythontips.com/en/latest/map_filter.html). Date accessed: Jan 12, 2020.

Answer More than half of the respondents i.e. 66% are from United States followed by United Kingdom (10%), Canada (3.6%), Australia (2.9%) and Sweden (1.3%)

11. What comments can you make about the United States, the United Kingdom and Germany? Are these results consistent with what you expected?
Explain why.

Answer More than half of the respondents are from United States, i.e. around 66%, which was quite expected. But those from United Kingdom is just 10% and that of Germany is even lower i.e. around 1% (similar to

Sweden) which were not expected as I expected better results ,they being the developed countries in the sector of IT.

> 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 top 5 countries.
> Calculate the mean, median and count for the ages of each gender for each of these countries.
>
> *Hint: You may need to create a copy or slice.*

In [682]:

```
# Your Code
df = dataset_a1
#list of the values of the top 5 countries
country_names = list(df['Country'].value_counts().head(5).keys())

#create new dataset of only the top 5 countries present in the list
country_dataset = df[df['Country'].isin(country_names)]

#function to calculate the mean,median,count on required columns
fun = {'Gender':{'Count':'count'},
       'Age':{'Mean':'mean','Median':'median'}}

#groupby Country and Gender and apply the agg function
groupbyGC = country_dataset.groupby(['Country',
                                     'Gender']).agg(fun).astype(int)
groupbyGC.columns = groupbyGC.columns.droplevel(0) # drop level 0 index
groupbyGC = groupbyGC.reset_index() # reset its index
groupbyGC
```

Out[682]:

```
         Country  Gender  Count  Mean  Median
0      Australia  Female     44    27      26
1      Australia    Male    123    26      25
2         Canada  Female     41    26      25
3         Canada    Male    169    26      25
4         Sweden  Female     20    27      25
5         Sweden    Male     57    26      25
6 United Kingdom  Female    164    25      24
7 United Kingdom    Male    420    24      22
8  United States  Female   1384    28      26
9  United States    Male   2462    26      23
```

@timgeb. (Oct 31,2018). Answer to question: Filter dataframe matching column values with list values in python [duplicate]. Retrieved from: https://stackoverflow.com/questions/53082014/filter-dataframe-matching-column-values-with-list-values-in-python (https://stackoverflow.com/questions/53082014/filter-dataframe-matching-column-values-with-list-values-in-python) . Date accessed: Jan 12, 2020.

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

Answer The respondents from all the countries have female count much lower than that of male i.e. even lower than half of the male count. But only in United States it is around half of that of males. The mean and median age of respondents from all the countries is around same i.e. 24-26 years of age.

1.4 Roles

Now let's investigate the different roles assumed by IT professionals and how they are distributed. Since we are specifically interested in data science, we will also create a flag for each of the participants to indicate whether his/her role is data-science related.

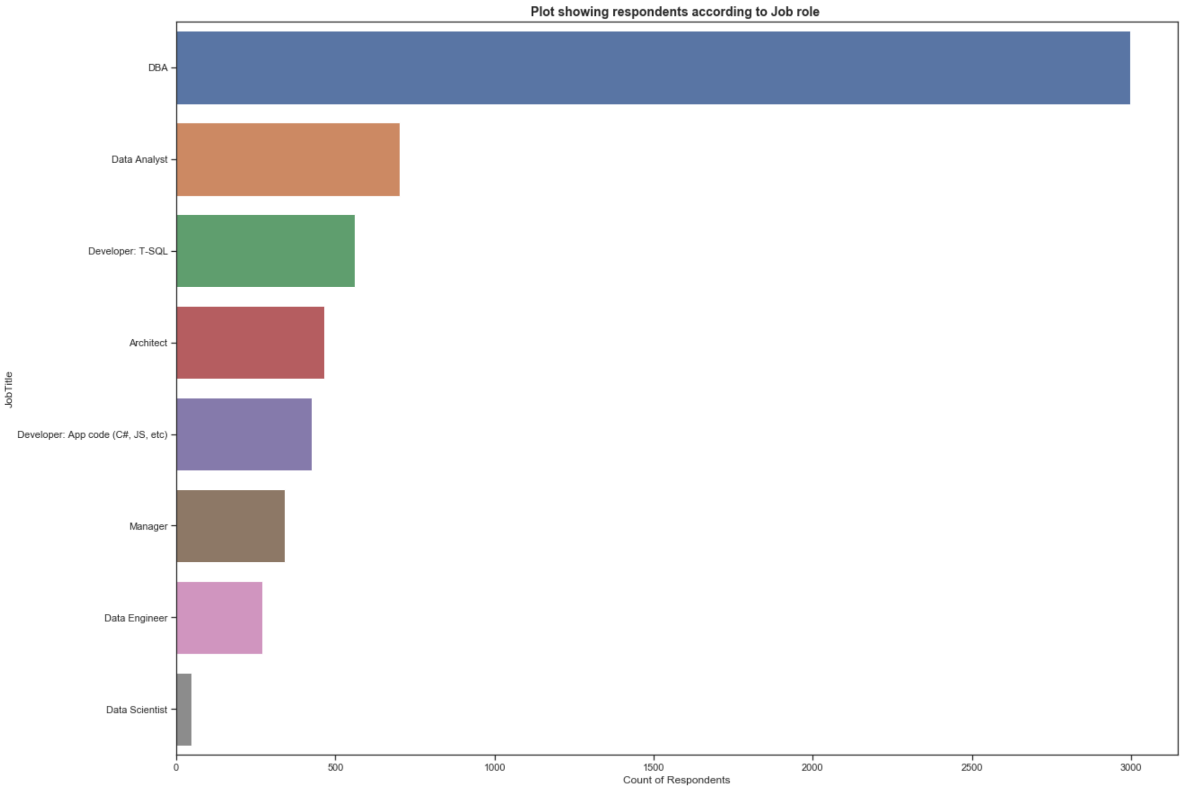
14. Plot a bar graph depicting the counts of different roles (each bar should represent the count of participants assuming a certain job role).

In [675]:

```
# Your code
plt.figure(figsize=(20,15)) #set the fig size

#plots the countplot with palette color blue
ax = sns.countplot(y="JobTitle", data = dataset_a1,
                   order=dataset_a1.JobTitle.value_counts().index)

#set the title in bold and particular fontsize and the axis labels
ax.set_title("Plot showing respondents according to Job role ",
             weight='bold').set_fontsize('14')
ax.set(xlabel = 'Count of Respondents')
```

Out[675]:

```
[Text(0.5, 0, 'Count of Respondents')]
```



15. What is the percentage of Data Scientists among the survey respondents?

In [611]:

```
# Your code
df = dataset_a1.JobTitle
jobtitle = "Data Scientist"

#match the data scientists and calculate their count
c_datascientist = df[df == jobtitle].count()
total = df.count() #total respondents count

#calculate the percentage using previously
#created perc function and print the result
print('Percent of Data Scientists:', perc(c_datascientist,total))
```

```
Percent of Data Scientists: 0.83
```

Answer Percentage of data scientists among the survey respondents is less than 1% i.e. 0.83%

16. Data Scientists usually work closely with specific functions in organisations. Data Analysts and Data Engineers are among the top collaborators with Data Scientists. Since our analysis will now focus on data science roles.

Create a boolean column "DataScienceRelated" which holds if a participant has a job title among "Data Scientist, Data Analyst or Data Engineer."

In [668]:

```
# Your code
df = dataset_a1
#create the set of required job titles
myset = {"Data Scientist", "Data Analyst" , "Data Engineer"}
#create a new boolean column which holds true if the
#job title matches with that in the myset
df['DataScienceRelated'] = df['JobTitle'].isin(myset)
#df
```

17. What is the percentage of Data Science related roles among the survey participants?

In [669]:

```
# Your code

df = dataset_a1.DataScienceRelated

#count of the respondents who has the data science related role
c_dsroles = df[df == True].count()
total = df.count() #total count

#calculate the percentage using previously
#created perc function and print the result
print('Percent of Data Science related roles:', perc(c_dsroles,total))
```

```
Percent of Data Science related roles: 17.56
```

Answer
The percentage of data science related roles is just 17.56 % of that of the total roles.

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 data scientists 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.

1. Plot a bar chart showing the percentage of each type of education for the three data science related roles.

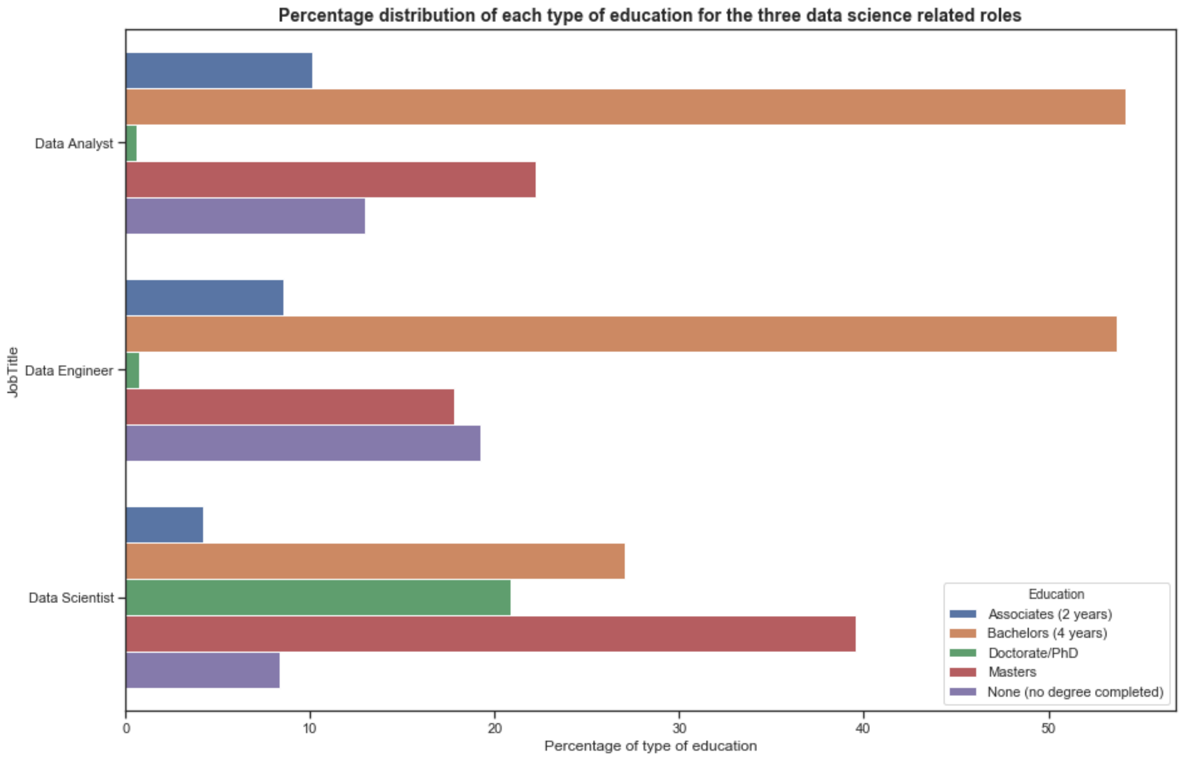
Hint: You should appropriately label your axes with a legend and a title

In [670]:

```
# Your code
df = dataset_a1

#create the new dataset for the data science related roles only
df_dsroles = df[df.DataScienceRelated == True]
#keep the columns JobTitle and Education
df_dsroles = df_dsroles[['JobTitle','Education']]

#create the column which has the total count of each job title
df_dsroles['Total_Count'] = df_dsroles.groupby([
                            'JobTitle']).transform('count')

#create the column which has the count of each eduction types in each job title
df_dsroles['Count'] = df_dsroles.groupby(['JobTitle','Education',
                     'Total_Count'])['Education'].transform('count')

#calculate the percentages and store it in a column called percent
df_dsroles['Percent'] = perc(df_dsroles['Count'],
                             df_dsroles['Total_Count'])

#create a new dataset which has the columns from the old grouped together
df_final = df_dsroles.groupby(['JobTitle','Education',
                               'Total_Count','Count','Percent']).count()
df_final = df_final.reset_index() # reset its index

#plot the  barplot
plt.figure(figsize=(15,10))  #set the fig size

#plot the barplot according to the jobs and education
ax = sns.barplot(y="JobTitle", x="Percent",
                 hue="Education" ,data = df_final)

#set the title in bold and particular fontsize and the axis labels
ax.set_title(
    "Percentage distribution of each type of education for the \
three data science related roles",
             weight='bold').set_fontsize('14')
ax.set(xlabel = 'Percentage of type of education')
```

Out[670]:

```
[Text(0.5, 0, 'Percentage of type of education')]
```



@EdChum. (Apr 22,2015). Answer to question: Python pandas: Add a column to my dataframe that counts a variable. Retrieved from: [https://stackoverflow.com/questions/29791785/python-pandas-add-a-column-to-my-dataframe-that-counts-a-variable (https://stackoverflow.com/questions/29791785/python-pandas-add-a-column-to-my-dataframe-that-counts-a-variable)](https://stackoverflow.com/questions/29791785/python-pandas-add-a-column-to-my-dataframe-that-counts-a-variable) . Date accessed: Jan 12, 2020.

2. 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 or is it job-dependent?

Answer According to the graph plotted, the Master's or Doctoral degree is not really taken into consideration for Data Analyst and Data Engineer roles. But for Data Scientist the respondents with Master's degree are more. This can be because the Data Scientist role is much regarding analysis and high level decision making of the business. Hence, an expert in the practical study or tools might be preferred for such a role with equal level of business understanding skills. But, also, requirements for any role in IT is purely job-dependent according to the global market situation and the job search sites like seek,linkedIn etc. with still some basic requirements necessary for the role.

3. Let's see if the trend is reflected in the Australian respondents.
Plot a bar chart like above but only for Australia, and display the counts of the number of Australian respondents holding a Doctoral degree for each of the three job roles as text output.

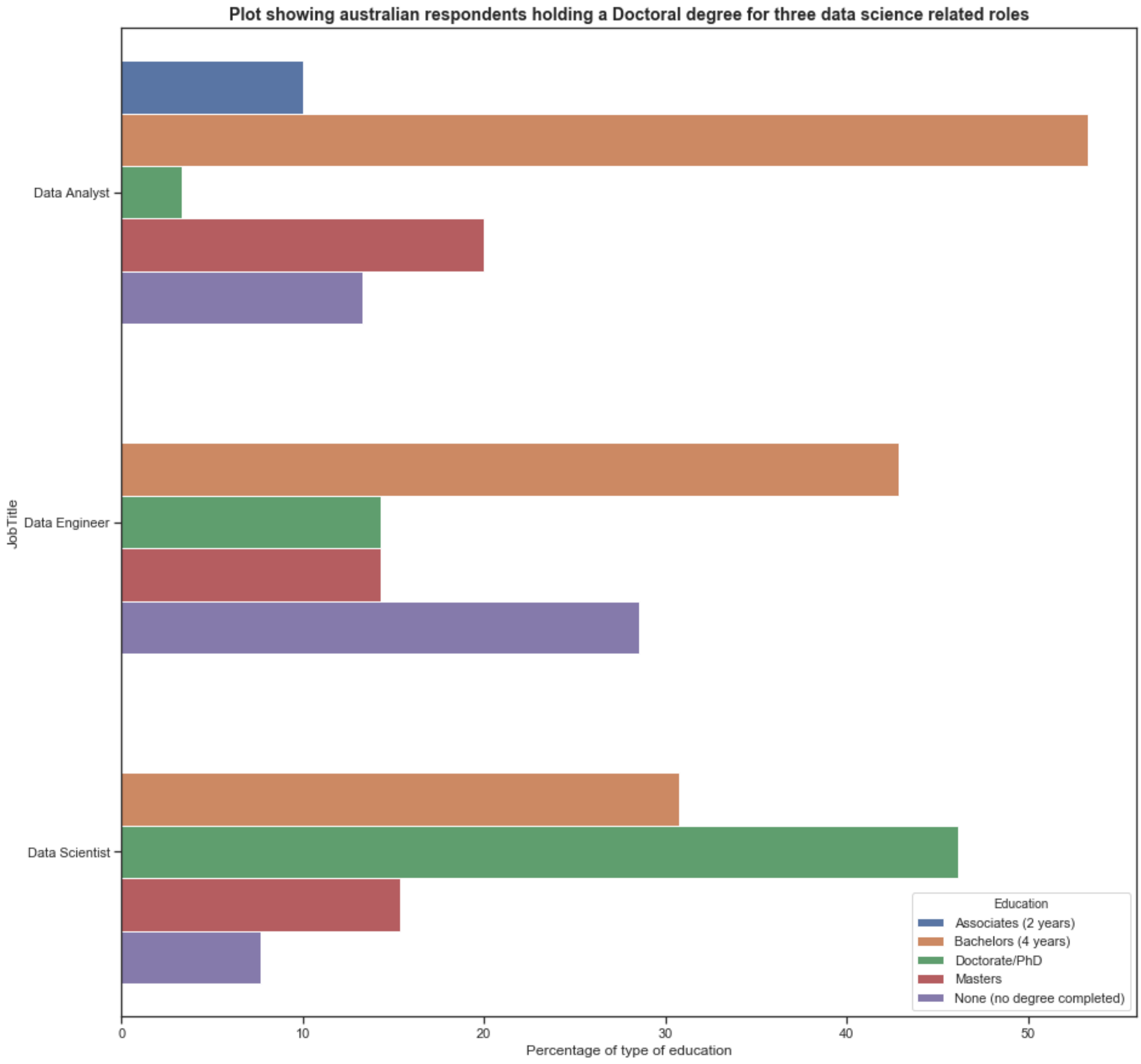
In [671]:

```
# Your code

plt.figure(figsize=(15,15)) #set the fig size
df = dataset_a1

#create the new dataset for the data science related roles and austalia only
df_dsroles = df[(df.DataScienceRelated == True) &
                (df.Country == 'Australia')]
#keep the columns JobTitle and Education
df_dsroles = df_dsroles[['JobTitle','Education']]

#create the column which has the total count of each job title
df_dsroles['Total_Count'] = df_dsroles.groupby([
                            'JobTitle']).transform('count')

#create the column which has the count of each eduction types in each job title
df_dsroles['Count'] = df_dsroles.groupby(['JobTitle','Education',
                    'Total_Count'])['Education'].transform('count')

#calculate the percentages and store it in a column called percent
df_dsroles['Percent'] = perc(df_dsroles['Count'],
                             df_dsroles['Total_Count'])

#create a new dataset which has the columns from the old grouped together
df_final = df_dsroles.groupby(['JobTitle','Education',
                               'Total_Count','Count',
                               'Percent']).count()
df_final = df_final.reset_index() # reset its index

#plot the  barplot

#plot the barplot according to the jobs and education
ax = sns.barplot(y="JobTitle", x="Percent",hue="Education" ,
                 data = df_final)

#set the title in bold and particular fontsize and the axis labels
ax.set_title(
    "Plot showing australian respondents holding a Doctoral degree \
for three data science related roles",
             weight='bold').set_fontsize('14')
ax.set(xlabel = 'Percentage of type of education')

#calculate the count of the australian respondents with PhD
#display in new dataframe- df_PhD
df_PhD = df_final[df_final.Education == 'Doctorate/PhD'][['JobTitle',
                                        'Education','Count']]
df_PhD.reset_index(drop=True, inplace=True)  # reset its index
df_PhD
```

Out[671]:

	JobTitle	Education	Count
0	Data Analyst	Doctorate/PhD	1
1	Data Engineer	Doctorate/PhD	1

	JobTitle	Education	Count
2	Data Scientist	Doctorate/PhD	6



@Subhojit Mukherjee. (Mar 6,2018). Answer to question:Removing index column in pandas when reading a csv. Retrieved from: [https://stackoverflow.com/questions/20107570/removing-index-column-in-pandas-when-reading-a-csv (https://stackoverflow.com/questions/20107570/removing-index-column-in-pandas-when-reading-a-csv)](https://stackoverflow.com/questions/20107570/removing-index-column-in-pandas-when-reading-a-csv) . Date accessed: Jan 15, 2020.

@Data to Fish. (Jan 6,2020). Answer to question: 5 ways to apply an IF condition in pandas DataFrame. Retrieved from: [https://datatofish.com/if-condition-in-pandas-dataframe/ (https://datatofish.com/if-condition-in-pandas-dataframe/)](https://datatofish.com/if-condition-in-pandas-dataframe/) . Date accessed: Jan 15, 2020.

4. Display as text output the mean and median age of ALL respondents according to each degree type.

In [620]:

```
# Your code
df = dataset_a1

#function to calculate the mean,median on Age
fun = {'Age':{'Mean Age':'mean','Median Age':'median'}}

#groupby Education and apply the agg function
groupby_education = df.groupby(['Education']).agg(fun).astype(int)
groupby_education = groupby_education.reset_index() # reset its index
# drop level 0 index
groupby_education.columns = groupby_education.columns.droplevel(0)
# rename the first column
groupby_education.rename(columns = {'':'Education'},inplace = True)
groupby_education
```

Out[620]:

	Education	Mean Age	Median Age
0	Associates (2 years)	26	24
1	Bachelors (4 years)	26	24
2	Doctorate/PhD	31	29
3	Masters	27	25
4	None (no degree completed)	26	24

3. Employment

Many of you will be seeking work after your degree. Let's have a look at the state of the employment market for the respondents of the survey.

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.

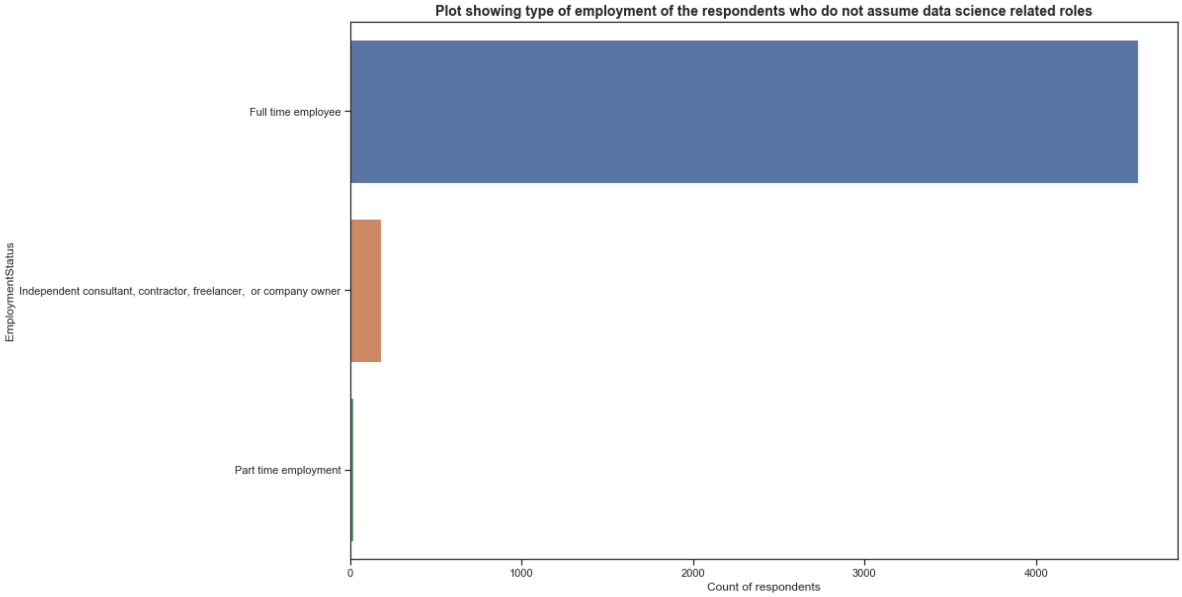
1. Plot the type of employment the respondents have on a bar chart for respondents who do not assume data science related roles.

In [672]:

```
# Your code
plt.figure(figsize=(15,10)) #set the fig size

#plots the countplot for EmploymentStatus
ax = sns.countplot(y="EmploymentStatus",
         data = dataset_a1[dataset_a1.DataScienceRelated == False] )

#set the title in bold and particular fontsize and the axis labels
ax.set_title(
    "Plot showing type of employment of the respondents \
who do not assume data science related roles",
             weight='bold').set_fontsize('14')
ax.set(xlabel = 'Count of respondents')
```

Out[672]:

```
[Text(0.5, 0, 'Count of respondents')]
```



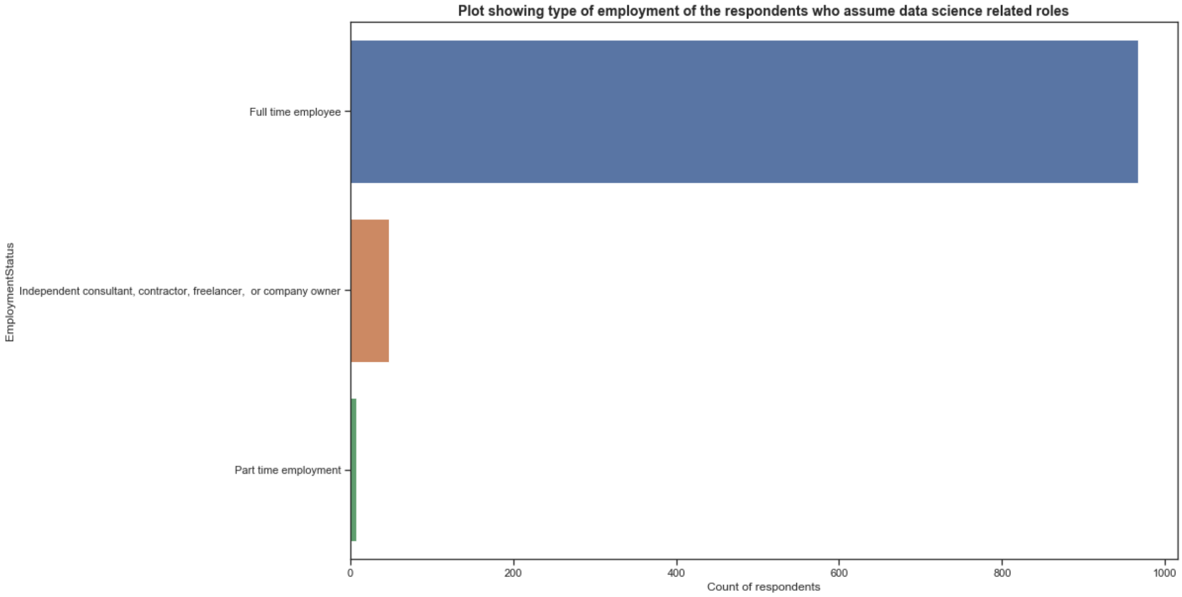
2. Now plot the type of employment the respondents have on a bar chart only for those assuming data science related roles

In [673]:

```
# Your code
plt.figure(figsize=(15,10)) #set the fig size

#plots the countplot for EmploymentStatus
ax = sns.countplot(y="EmploymentStatus",
         data = dataset_a1[dataset_a1.DataScienceRelated == True] )

#set the title in bold and particular fontsize and the axis labels
ax.set_title("Plot showing type of employment of the respondents \
who assume data science related roles",
             weight='bold').set_fontsize('14')
ax.set(xlabel = 'Count of respondents')
```

Out[673]:

```
[Text(0.5, 0, 'Count of respondents')]
```



3. Comparing the two graphs, would you say that the data science roles differ in the type of employment as opposed to non-data science roles?
Explain your answers.

Answer Both the graphs, for data science related roles and non data science related roles shows same ratio as compared among the full time, independent employment and part time. This is definitely because most of the requirements of the IT jobs are for full time and are paid well and also require sufficient amount of efforts/time consuming efforts.Also, privacy and security of the projects and clients play an important role which cannot be ignored. Hence, the companies would indulge into contracts etc. to prevent the information leakage and to bring stability for IT governance. Also, on a commercial basis, it is easy to get the licences for the advanced technologies available in the market which is not possible on an individual level.

Kapoor,N.(2017).Freelancing vs Full-time Job: Pros & Cons.Retrieved from https://yourstory.com/mystory/fa474e2c2d-freelancing-vs-full-ti (https://yourstory.com/mystory/fa474e2c2d-freelancing-vs-full-ti)

4. Let's investigate whether the type of employment is country dependent.
Print out the percentages of all respondents who are employed full time in Australia, United Kingdom and the United States.

In [679]:

```
# Your code
df = dataset_a1

#create the set of required countries
myset = {"Australia", "United Kingdom" , "United States"}

#create the new dataset for the countries in the set only
df_modified= df[df.Country.isin(myset)]
#keep the columns Country and Employmentstatus
df_modified = df_modified[['Country','EmploymentStatus']]

#create the column which has the total count of each country
df_modified['Total_Count'] = df_modified.groupby(
                            ['Country']).transform('count')

#create new dataset for full time employees only
#copying the columns from old dataset (data and indices both)
df = df_modified[df_modified.EmploymentStatus ==
                 "Full time employee"].copy(deep=True)

#create the column which has the count of each Country and Employmentstatus
df['Count'] = df.groupby(['Country',
                'EmploymentStatus',
                'Total_Count'])['EmploymentStatus'].transform('count')

#calculate the percentages and store it in a column called percent
df['Percent'] = perc(df['Count'],df['Total_Count'])

#create a new dataset which has the columns from the old grouped together
df_final = df.groupby(['Country','EmploymentStatus',
                       'Total_Count','Count','Percent']).count()
df_final = df_final.reset_index() #reset index
df_final = df_final.drop(["EmploymentStatus","Total_Count","Count"],axis = 1)
df_final
```

Out[679]:

	Country	Percent
0	Australia	87.43
1	United Kingdom	92.64
2	United States	97.63

@cs95. (Dec 28,2018). Answer to question:How to deal with SettingWithCopyWarning in Pandas?. Retrieved from: https://stackoverflow.com/questions/20625582/how-to-deal-with-settingwithcopywarning-in-pandas (https://stackoverflow.com/questions/20625582/how-to-deal-with-settingwithcopywarning-in-pandas) . Date accessed: Jan 16, 2020.

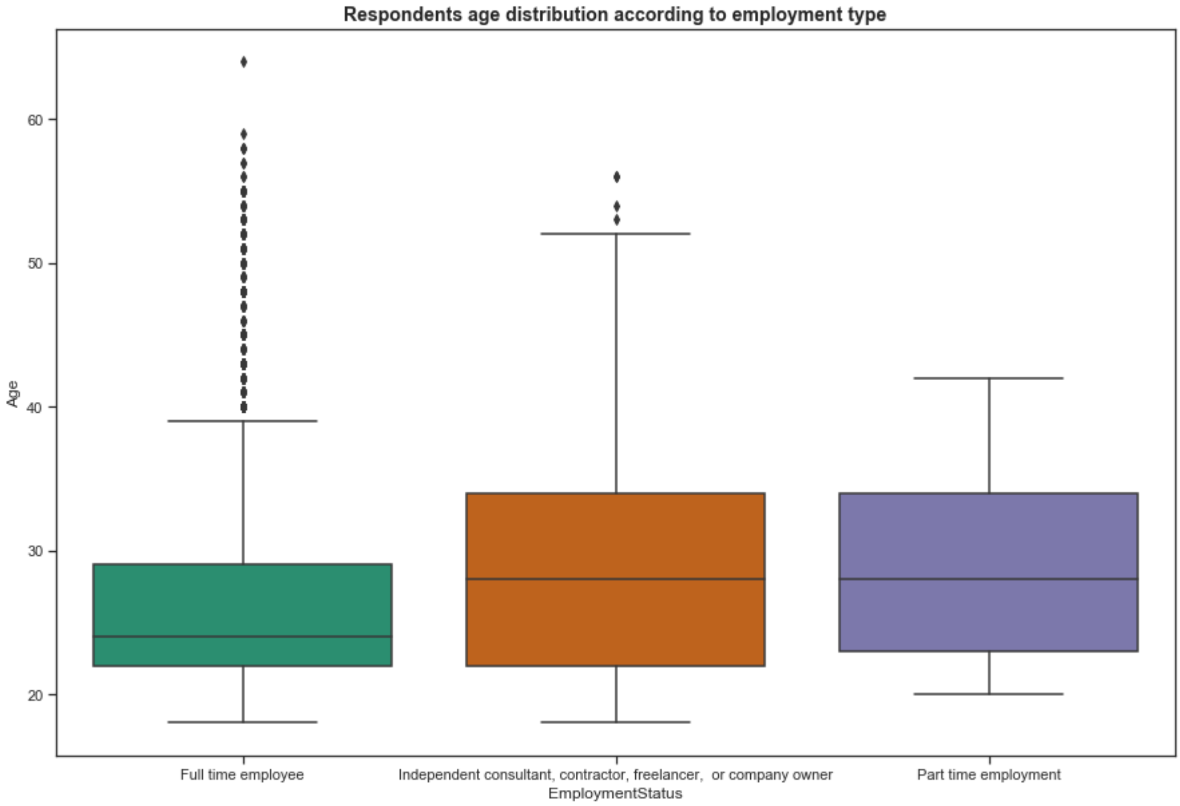
Remember earlier, we saw that age seemed to have some interesting characteristics when plotted with other variables.

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

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

In [626]:

```
# Your code
plt.figure(figsize=(15,10)) #set the fig size
#plots the boxplot for EmploymentStatus and Age
ax = sns.boxplot(x="EmploymentStatus",y="Age",
                 data=dataset_a1,palette = 'Dark2')

#set the title in bold and particular fontsize and the axis labels
ax.set_title(
    "Respondents age distribution according to employment type",
             weight='bold').set_fontsize('14')
```



6. What are your observations?

Answer The maximum age of the respondents is less for full time employees as compared to that of the part time and independent employees. The maximum age is highest for that of the independent employees. The median age of respondents is lower than others for the full time employees and positively skewed. While the other plots are uniformly skewed from first to third quartile and has the median age equal but higher than that of the fulltime employees. Hence, it can be observed that independent and part time employment is opted by most IT professionals not mostly in the start of their career. And, is carried out till a higher age as compared to the full time employees.

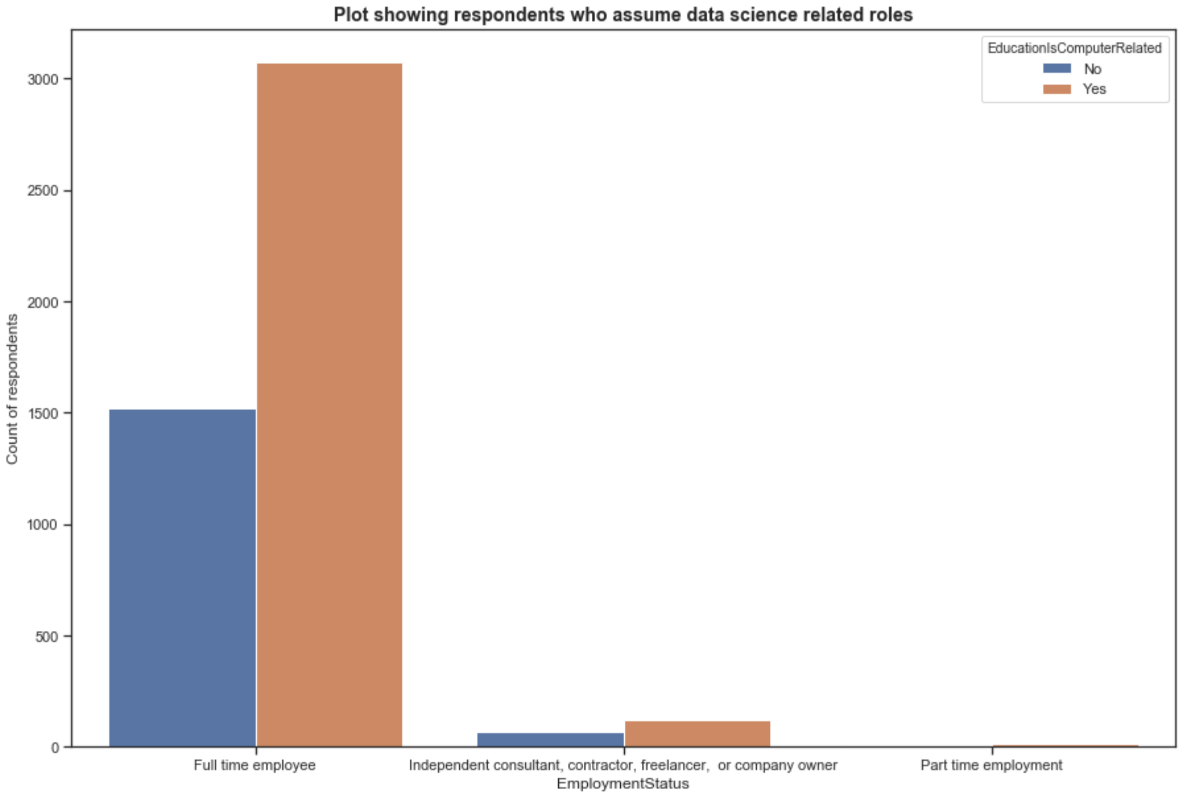
> 7. You may be wondering if a relevant Computer degree is necessary to help gain full-time employment after graduation.
> Plot the respondents' employment types (for all respondents) for each of the two categories of "EducationIsComputerRelated".

In [306]:

```
# Your code
plt.figure(figsize=(15,10)) #set the fig size
#plots the countplot for EmploymentStatus and EducationIsComputerRelated
ax = sns.countplot(x="EmploymentStatus",
                   hue="EducationIsComputerRelated" ,data = dataset_a1 )

#set the title in bold and particular fontsize and the axis labels
ax.set_title("Plot showing respondents who assume data science related roles",
             weight='bold').set_fontsize('14')
ax.set(ylabel = 'Count of respondents')
```

Out[306]:

```
[Text(0, 0.5, 'Count of respondents')]
```



> 8. Looking at the graph, does holding a computer-related degree improves your chances of securing a full-

time job?
Explain your answers.

Answer According to the graph, holding a computer related degree does improve the chances of securing a full time job as well for independent and part time employment though their total number is very less. We can say that it does improve the chances by 50%, which is a very significant number. We may predict that some computer relate degree assures the hirers that the employee would be familiar with the IT culture and would have atleast some basic knowledge from where he can catch up on to the required technical knowledge. Hence, the chances of landing in a full time job increases.

3.2 Job Satisfaction

Let's now investigate how happy IT professionals are about their jobs. It is also relevant to look at the years of experience to see whether the job gets boring after a while.

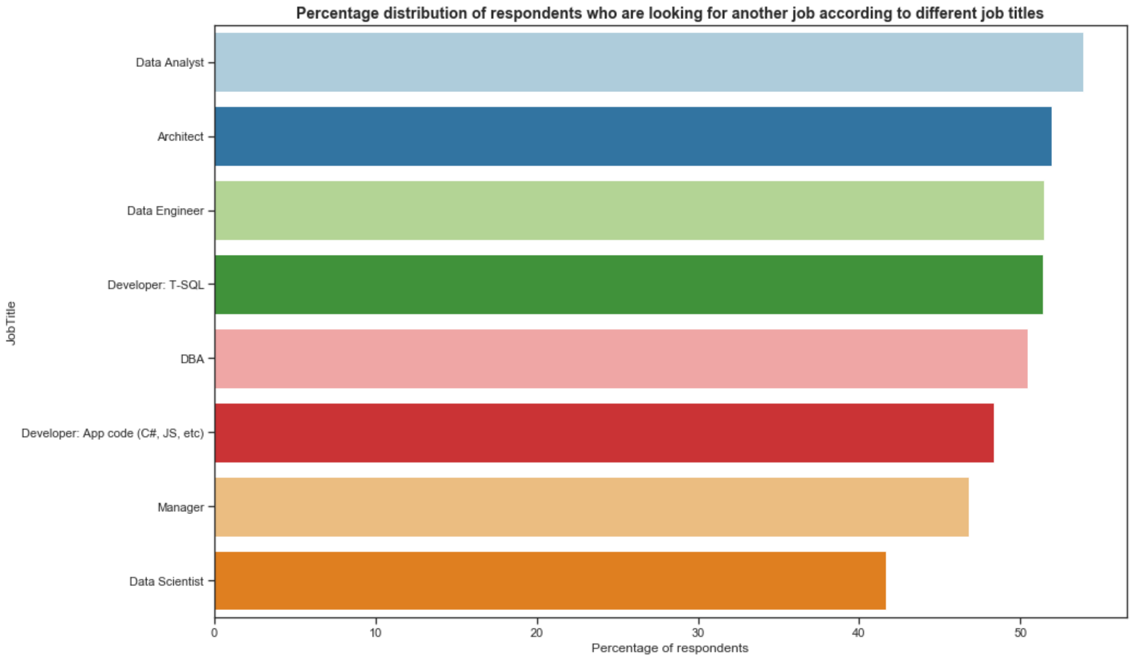
9. Create a bar chart for the percentage of respondents who are looking for another job grouped by the different job titles.

In [646]:

```
# Your code
plt.figure(figsize=(15,10)) #set the fig size
df_org = dataset_a1

#create new dataset by copying the columns  JobTitle and
#LookingForAnotherJob from old dataset (data and indices both)
df_job = df_org[['JobTitle','LookingForAnotherJob']].copy(deep=True)

#create the column which has the total count of each Job
df_job['Total_Count'] = df_job.groupby(['JobTitle']).transform('count')

#create new dataset for respondents looking for other job only
#copying the columns from old dataset (data and indices both)
df = df_job[df_job.LookingForAnotherJob == 'Yes'].copy(deep=True)

#create the column which has the count of each
#JobTitle and LookingForAnotherJob
df['Count'] = df.groupby(['JobTitle',
            'LookingForAnotherJob',
            'Total_Count'])['LookingForAnotherJob'].transform('count')

#calculate the percentages and store it in a column called percent
df['Percent'] = perc(df['Count'],df['Total_Count'])

#create a new dataset which has the columns from the old grouped together
df_final = df.groupby(['JobTitle','LookingForAnotherJob',
                       'Total_Count','Count','Percent']).count()
df_final = df_final.reset_index() #reset index
df_final = df_final.sort_values("Percent",ascending = False)

#plots the barplot for Percent and JobTitle
ax = sns.barplot(x="Percent", y="JobTitle" ,
                 data = df_final, palette = 'Paired')

#set the title in bold and particular fontsize and the axis labels
ax.set_title(
"Percentage distribution of respondents who are looking for \
another job according to different job titles",
             weight='bold').set_fontsize('14')
ax.set(xlabel = 'Percentage of respondents')
```

Out[646]:

```
[Text(0.5, 0, 'Percentage of respondents')]
```



10. What are the two roles that have the highest and lowest percentage of employees looking for other jobs?

Answer The two roles that have the highest percentage of employees looking for other jobs are Data Analyst and Architect and two roles that have the lowest percentage of employees looking for other jobs are Data Scientist and Manager.

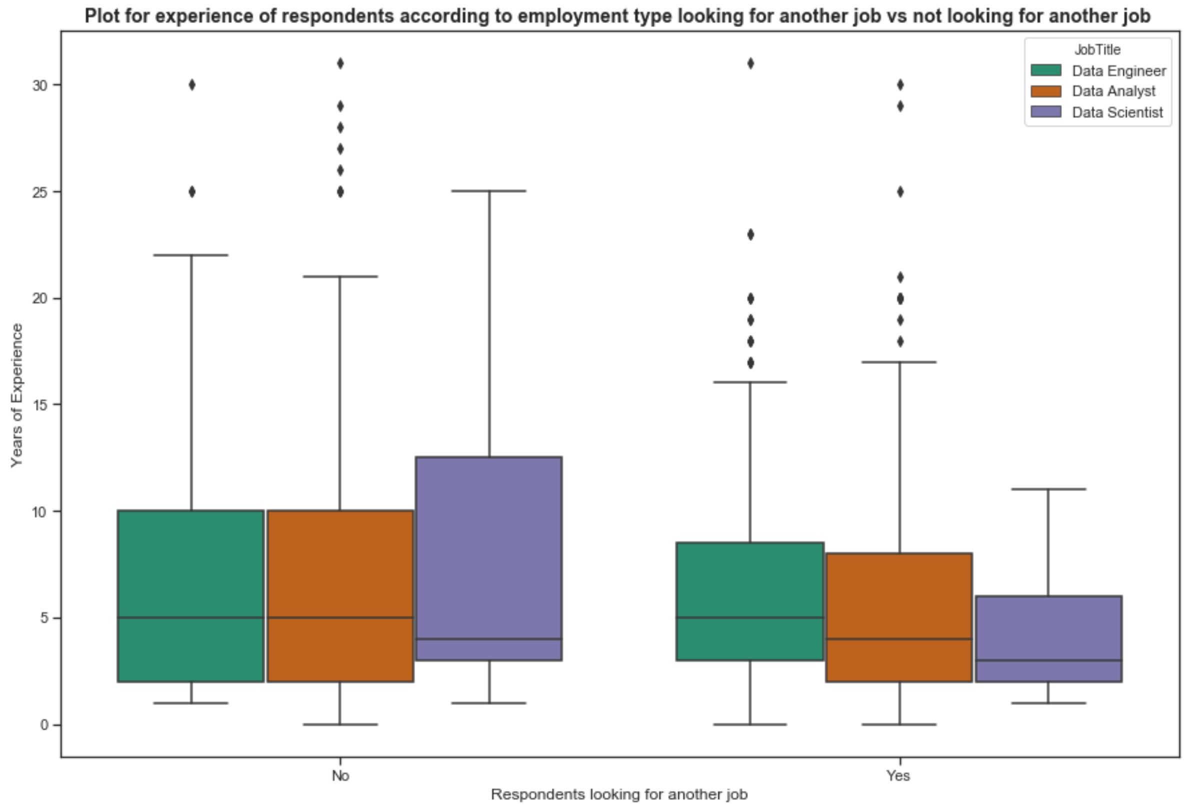
11. Let's focus on data science-related roles. Plot a box plot depicting the distribution of years-of-experience of those respondents who are looking for another job versus those who are not for each of the three roles.

In [628]:

```
# Your code
plt.figure(figsize=(15,10)) #set the fig size

#plots the boxplot for LookingForAnotherJob and YearsofExperience
ax = sns.boxplot(x="LookingForAnotherJob",y="YearsofExperience",
                 hue="JobTitle",
                 data = dataset_a1[dataset_a1.DataScienceRelated == True],
                 palette = 'Dark2')

#set the title in bold and particular fontsize and the axis labels
ax.set_title(
"Plot for experience of respondents according to employment type \
looking for another job vs not looking for another job ",
             weight='bold').set_fontsize('14')
ax.set(xlabel = 'Respondents looking for another job',
       ylabel = 'Years of Experience')
```

Out[628]:

```
[Text(0, 0.5, 'Years of Experience'),
 Text(0.5, 0, 'Respondents looking for another job')]
```



Plot for experience of respondents according to employment type looking for another job vs not looking for another job

12. What can you say about the years of experience as to whether it impacts happiness?

Answer From the boxplot, it is clear that the respondents looking for another job have low years as experience as compared with that of the respondents who are not looking for another job. Hence, it can be predicted that the higher the years of experience, the lower the chances that they would be in search of another job and be satisfied and happy with their current job. This can be explained as, an employee who is satisfied with the job profile and job role would be happy with it and opt to continue. But also,this can go other way, after high years of experience the employees would prefer adjusting and continuing in the same familiar environment instead of struggling to fit in a new one.

4. Salary

Data science is considered a very well paying role and was named 'best job of the year' for 2019.

We would like to investigate in this section the different salary ranges for the different job roles in the IT industry and compare it to those of Data Science roles.

4.1 Salary overview

Note that the salaries given in the dataset is in USD. If we are to investigate the salaries in AUD, we need to consider the currency conversion.

You can use the following rate of conversion:

```
1 USD = 1.47 AUD
```

Let's have a look at the data.

1. Create a derived column "SalaryAUD" containing the converted salary data into Australian Dollars (AUD).
Print out the maximum and median salary in AUD for each of the job roles in our dataset.

In [657]:

```python
# Your code
df = dataset_a1
#create a new column for Salary in AUD
#by applying the conversion factor given
df['SalaryAUD'] = df.apply(lambda x: x.SalaryUSD * 1.47, axis = 1)

#function to calculate the max,median on required column
fun = {'SalaryAUD':{'Median Salary':'median','Max Salary':'max'}}

#groupby JobTitle and apply the agg function
groupby_jobs = df.groupby(['JobTitle']).agg(fun).astype(int)
# drop level 0 index
groupby_jobs.columns = groupby_jobs.columns.droplevel(0)
groupby_jobs = groupby_jobs.reset_index() # reset its index
groupby_jobs
```

Out[657]:

	JobTitle	Median Salary	Max Salary
0	Architect	176400	514500
1	DBA	132300	1411200
2	Data Analyst	113190	624750
3	Data Engineer	139650	955500
4	Data Scientist	163170	235200
5	Developer: App code (C#, JS, etc)	117600	285180
6	Developer: T-SQL	124950	1036350
7	Manager	161700	924419

2. Do those figures confirm that data scientists are well paid?

Answer Though the median salary of a data scientist show a satisfactory number of 163170 AUD, the maximum salary is only $235200 AUD. This is pretty low as compared to the other job roles in IT. Hence, the figures fail to convince that data scientists are well paid.

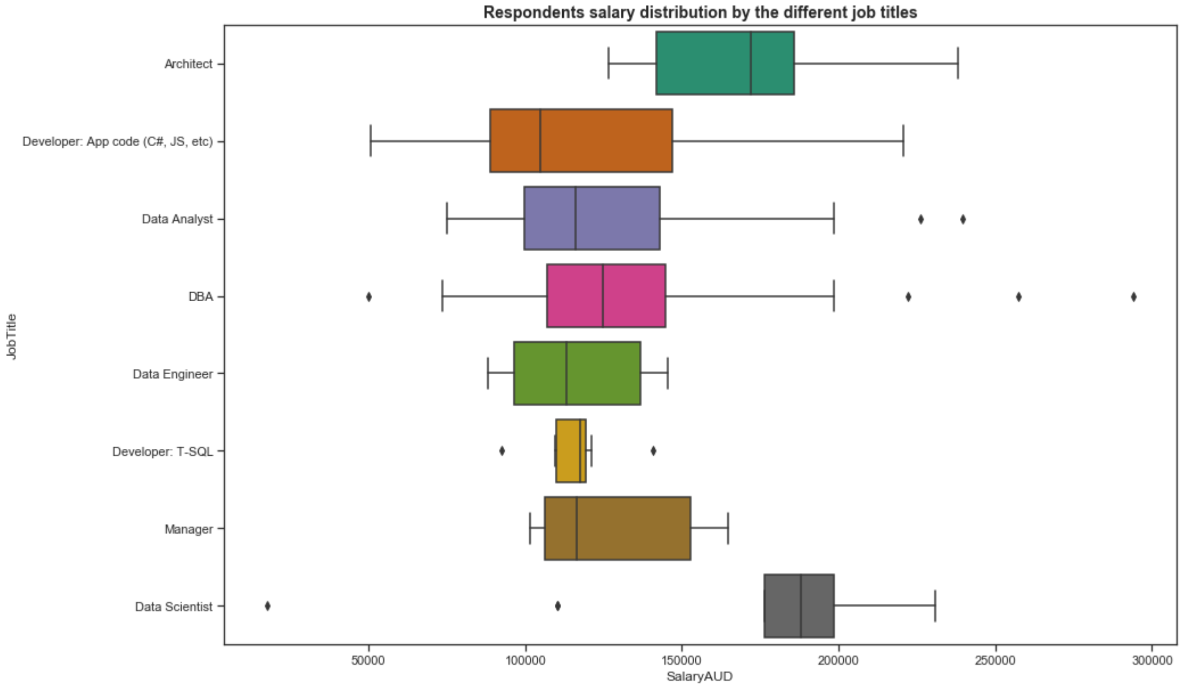
4.2 Salary by country

Since each country has different cost of living and pay indexes, we want to compare these jobs only in Australia.

3. Plot boxplot chart of the Australian respondents salary distribution grouped by the different job titles.

In [658]:

```
# Your code
plt.figure(figsize=(15,10)) #set the fig size

#plots the boxplot for JobTitleand SalaryAUD
ax = sns.boxplot(y="JobTitle",x="SalaryAUD",
                 data = dataset_a1[dataset_a1.Country == 'Australia'],
                 palette = 'Dark2')
#set the title in bold and particular fontsize and the axis labels
ax.set_title(
    "Respondents salary distribution by the different job titles",
             weight='bold').set_fontsize('14')
```



4. How are data scientists paid in comparison to other roles in Australia?

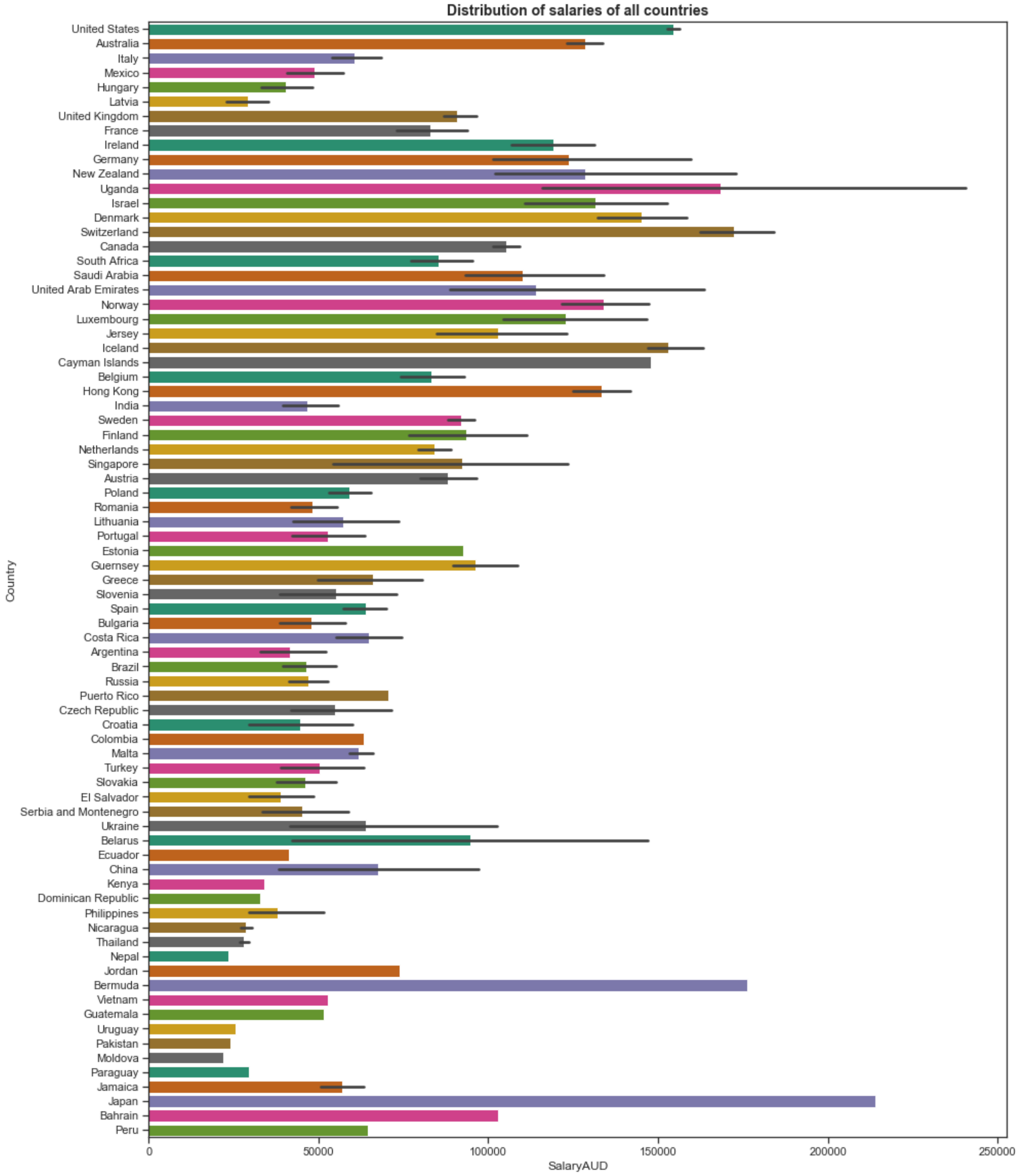
Answer Data Scientists are paid fairly high in Australia as compared to the other roles. The boxplot of the data scientist shows almost uniform skewness from first quartile to third quartile. Hence, the salaries seem to be almost evenly distributed around the median salary. The minimun salary is equal to the first quartile, hence the starting salary seems to be pretty good for data scientist in Australia. But the maximum salary does not show high difference from the median salary. Hence, the scope for the maximum salary achieveable in this field does not show much difference to that of the median salary.

5. Australia's salaries look pretty good in general. Is that the case for all other countries? Plot the salaries of all countries on a bar chart (with error bars).

Hint: Consider all job titles and filter for full-time employees only

In [642]:

```
# Your code
plt.figure(figsize=(15,20)) #set the fig size
#plots the barplot for Country and SalaryAUD
ax = sns.barplot(y="Country",x="SalaryAUD",
         data = dataset_a1[dataset_a1.EmploymentStatus ==
                           'Full time employee'],
         palette = 'Dark2')
#set the title in bold and particular fontsize and the axis labels
ax.set_title("Distribution of salaries of all countries",
             weight='bold').set_fontsize('14')
```



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

Answer The highest salary shows for Japan, Bermuda and Uganda but these numbers are misleading beacause there is only one respondent from Japan and Bermuda for the survey and only four respondents from Uganda. Hence the salary distribution on such a limited data cannot be trusted as accurate as compared to that to the United States , Australia etc. from where the respondents are high in number.

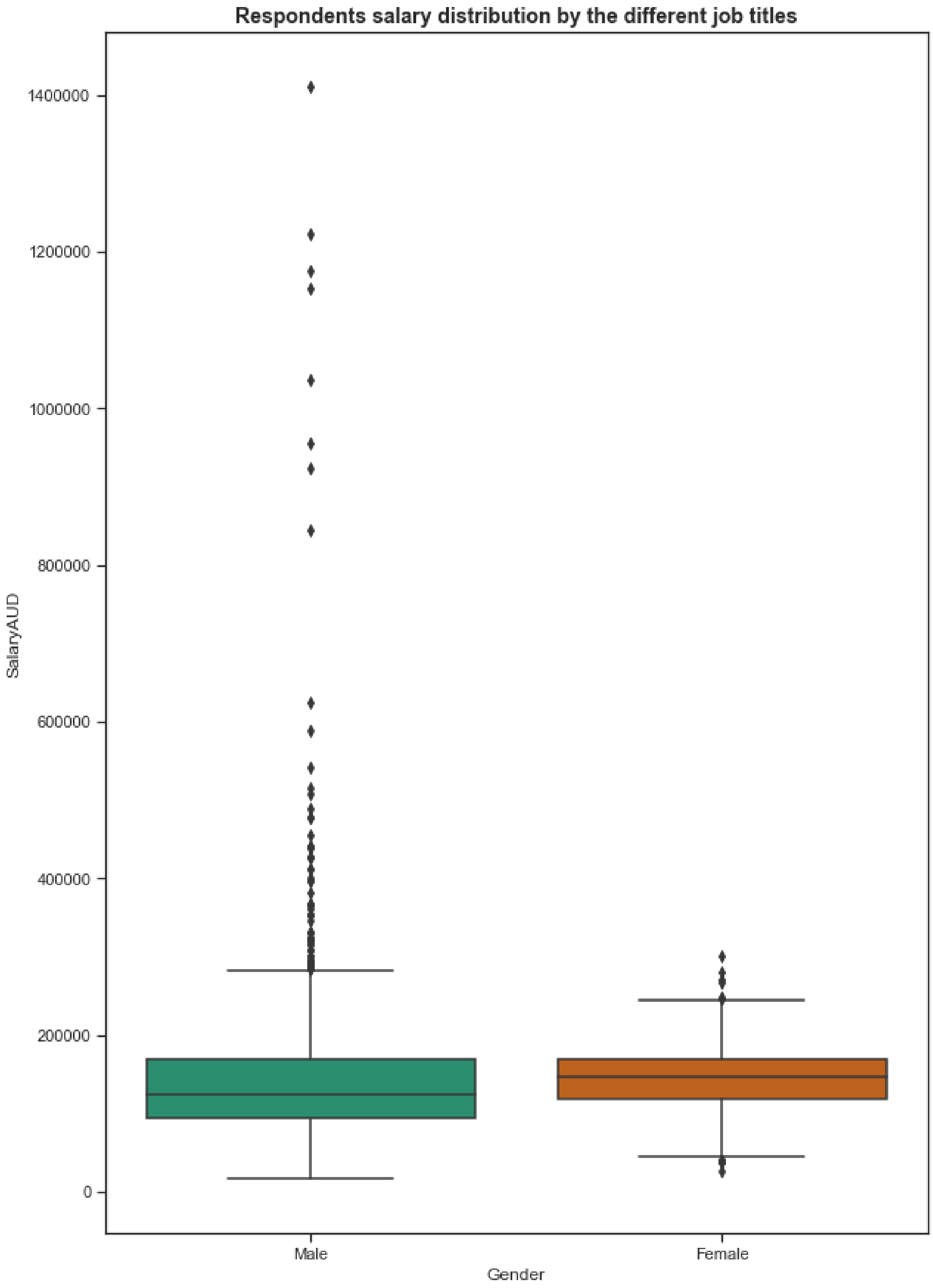
4.3 Salary and Gender

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

7. Plot the salaries of all respondents grouped by gender on a boxplot.

In [641]:

```
# Your code
plt.figure(figsize=(10,15)) #set the fig size
#plots the boxplot for Gender and SalaryAUD
ax = sns.boxplot(x="Gender",y="SalaryAUD",data = dataset_a1,
                 palette = 'Dark2')
#set the title in bold and particular fontsize and the axis labels
ax.set_title(
    "Respondents salary distribution by the different job titles",
             weight='bold').set_fontsize('14')
```



8. What do you notice about the distributions?

Answer The distribution shows that the boxplot for female is uniformly skewed on both ends. The median of the female is higher than that of male. But, the maximum salary of female is lower tahn that of the male. The minimum salary of female is higher than that of male. Hence, it can be observed that, the females are paid a higher starting salary as compared to that of males, but the highest salary achieved is less as compared to male. If we consider the outliers for male, the highest salary is significantly less for female. We can also predict that the top level job roles having significantly high salaries have more male employees than female employees.

Catalyst, Quick Take: Women in Management.(August 7, 2019).Retrieved from https://www.catalyst.org/research/women-in-management/ (https://www.catalyst.org/research/women-in-management/)

9. The salaries may be affected by the country the respondent 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 [660]:

```
# Your code
df = dataset_a1

#create the set of required countries
myset = {"Australia", "United States" , "India"}

#create the new dataset for the countries in the set only
df_countries = df[df['Country'].isin(myset)]

#function to calculate the median on required column
fun = {'SalaryAUD':{'Median Salary':'median'}}
#groupby 'Country','Gender' and apply the agg function
groupby_gender = df_countries.groupby(['Country',
                            'Gender']).agg(fun).astype(int)
# drop level 0 index
groupby_gender.columns = groupby_gender.columns.droplevel(0)
groupby_gender = groupby_gender.reset_index() # reset its index
groupby_gender
```

Out[660]:

	Country	Gender	Median Salary
0	Australia	Female	139650
1	Australia	Male	122010
2	India	Female	48142
3	India	Male	34251
4	United States	Female	147602
5	United States	Male	154350

4.4 Salary and formal education

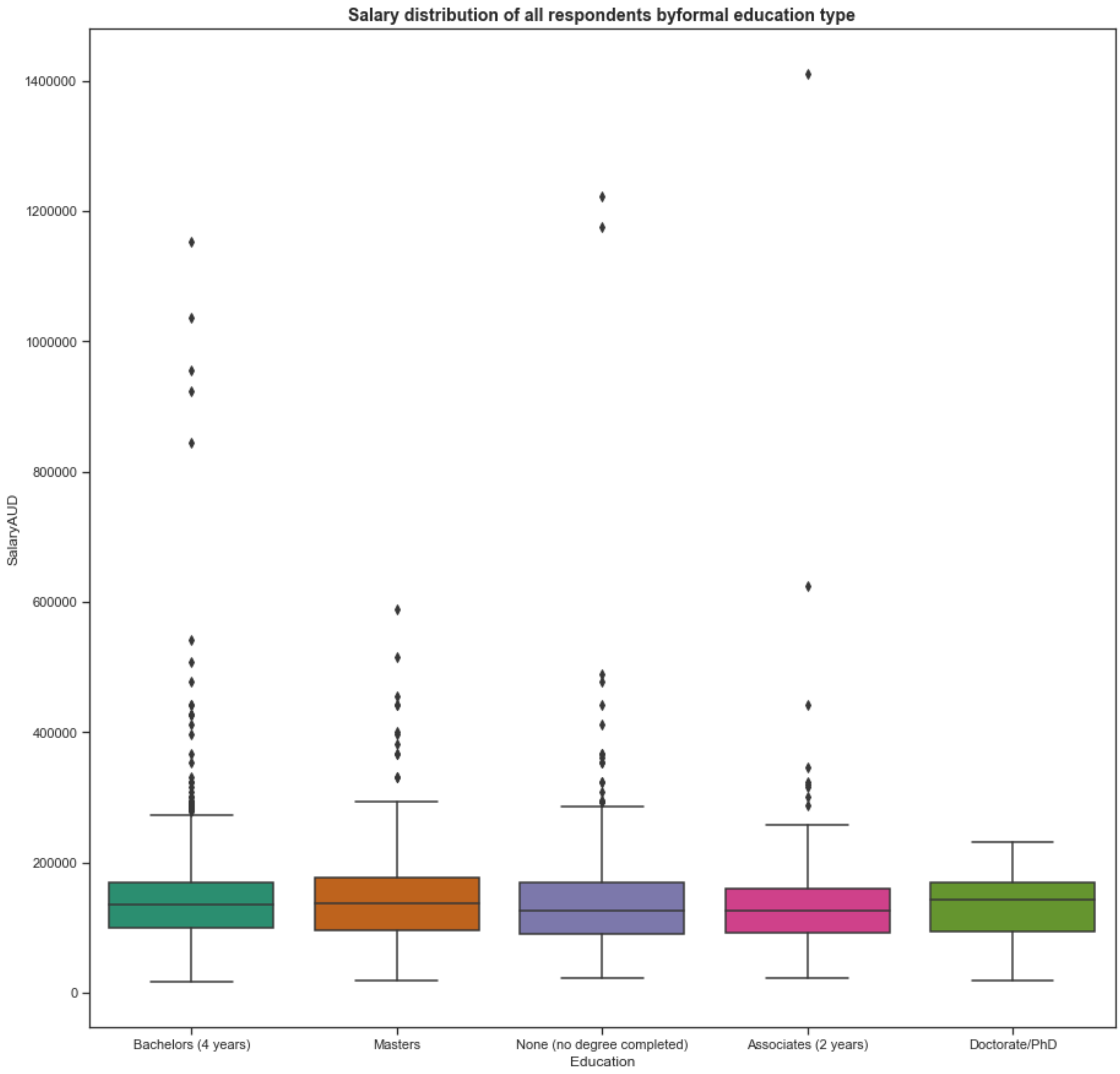
Is getting your master's really worth it ? Do PhDs get more money?

Let's see.

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

In [661]:

```
# Your code
plt.figure(figsize=(15,15)) #set the fig size
#plots the boxplot for Education and SalaryAUD
ax = sns.boxplot(x="Education",y="SalaryAUD",
                 data = dataset_a1,
                 palette = 'Dark2')
#set the title in bold and particular fontsize and the axis labels
ax.set_title("Salary distribution of all respondents by\
formal education type", weight='bold').set_fontsize('14')
```



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

Answer From the distribution it is clear that the maximum salary if highest for the Masters and also has many outliers as compared to the PhD. The boxplot of Masters is uniformly skewed between the first and the third quartile. The boxplot for PhD is left skewed. Hence more than 50% of the PhD respondenets are below the median salary of the PhDs. Also, it can be derive that the starting salary for the Masters is slightly higher than that of the PhD respondents. Hence, considering all the observations we can conclude that it is better to get a Masters degree as compared to the PhD.

4.5 Salary and Employment Sector

Do government jobs pay better than private sector? Does it differ based on the country?

Let's see.

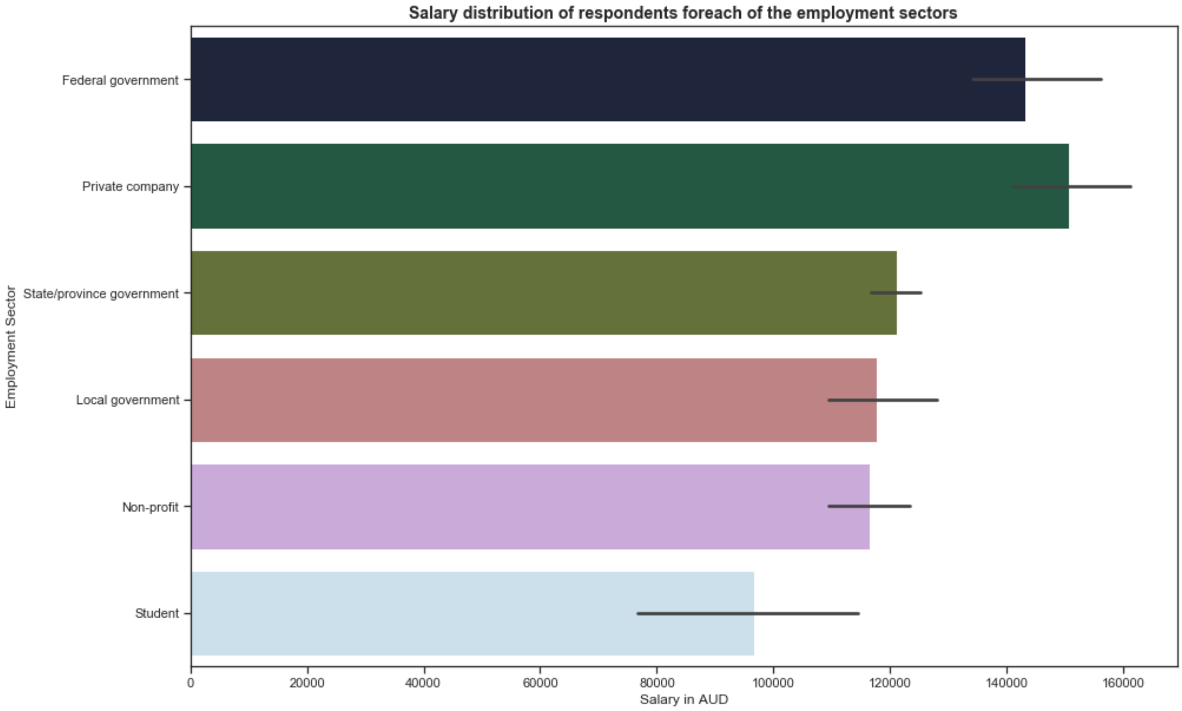
12. Plot a bar chart (with error bars) of the salaries of respondents for each of the employment sectors.

In [633]:

```
# Your code
plt.figure(figsize=(15,10)) #set the fig size
#plots the barplot for SalaryAUD and EmploymentSector
ax = sns.barplot(y="EmploymentSector",x="SalaryAUD",
                 data = dataset_a1,
                 palette = 'cubehelix')
#set the title in bold and particular fontsize and the axis labels
ax.set_title("Salary distribution of respondents for\
each of the employment sectors",
             weight='bold').set_fontsize('14')
ax.set(xlabel= 'Salary in AUD',ylabel = 'Employment Sector')
```

Out[633]:

```
[Text(0, 0.5, 'Employment Sector'), Text(0.5, 0, 'Salary in AUD')]
```



13. Which seems to be the highest paying sector overall?
Do you think it would differ based on the country?

Propose a method to find out and explain your answer.

Answer The private sectors seems to be the highest paying sector overall. But, this might differ based on the country. This is because in some countries the federal government can be the one paying highest salaries or may have the private companies to work in tie-ups with the government projects. There are various methods to find this out using the same dataset like, plotting a barplot graph by grouping the employment sectors according to the countries, by getting a numerical data for the percentage of employment sectors paying the highest salary for the countries.

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; one method we could use is a linear regression. This is a basic but powerful 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.

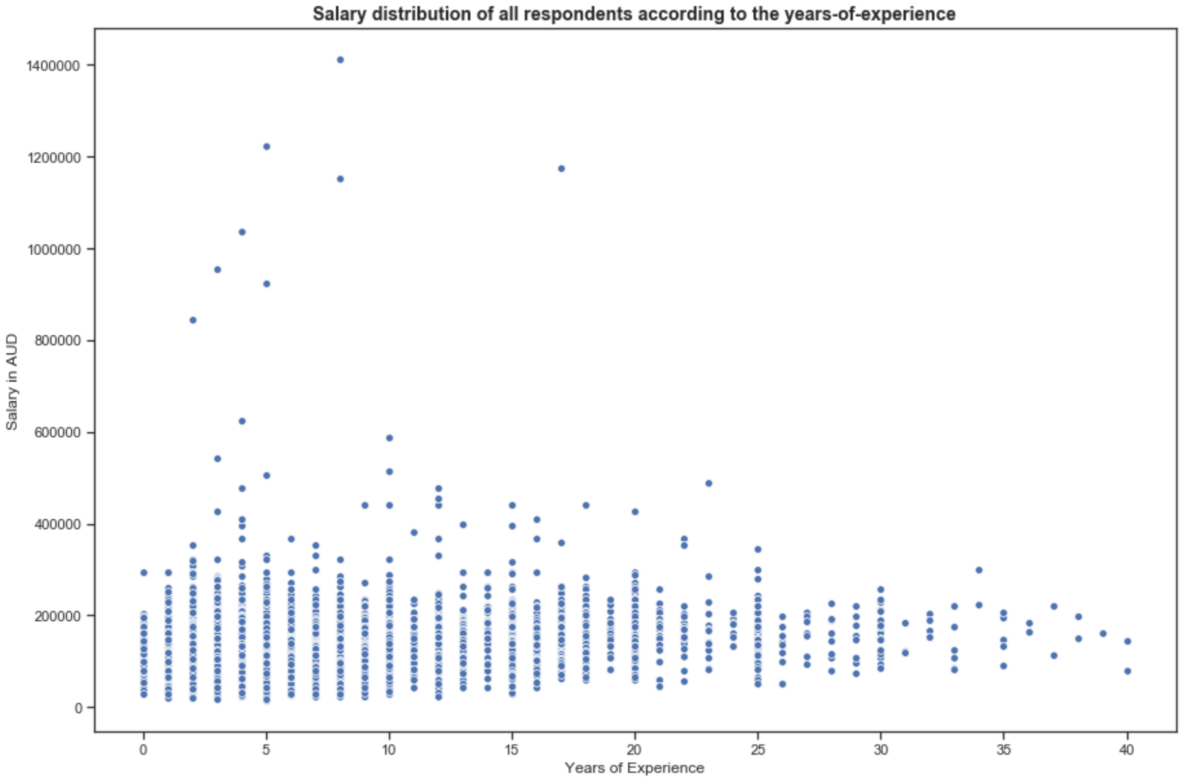
1. Plot the salary and years-of-experience of respondants on a scatterplot.

In [662]:

```
# Your code
plt.figure(figsize=(15,10)) #set the fig size
#plots the scatterplot for SalaryAUD and YearsofExperience
ax = sns.scatterplot(x="YearsofExperience",y="SalaryAUD",
                 data = dataset_a1,
                 palette = 'Dark2')
#set the title in bold and particular fontsize and the axis labels
ax.set_title(
"Salary distribution of all respondents according \
to the years-of-experience",
             weight='bold').set_fontsize('14')
ax.set(ylabel= 'Salary in AUD',xlabel = 'Years of Experience')
```

Out[662]:

```
[Text(0, 0.5, 'Salary in AUD'), Text(0.5, 0, 'Years of Experience')]
```



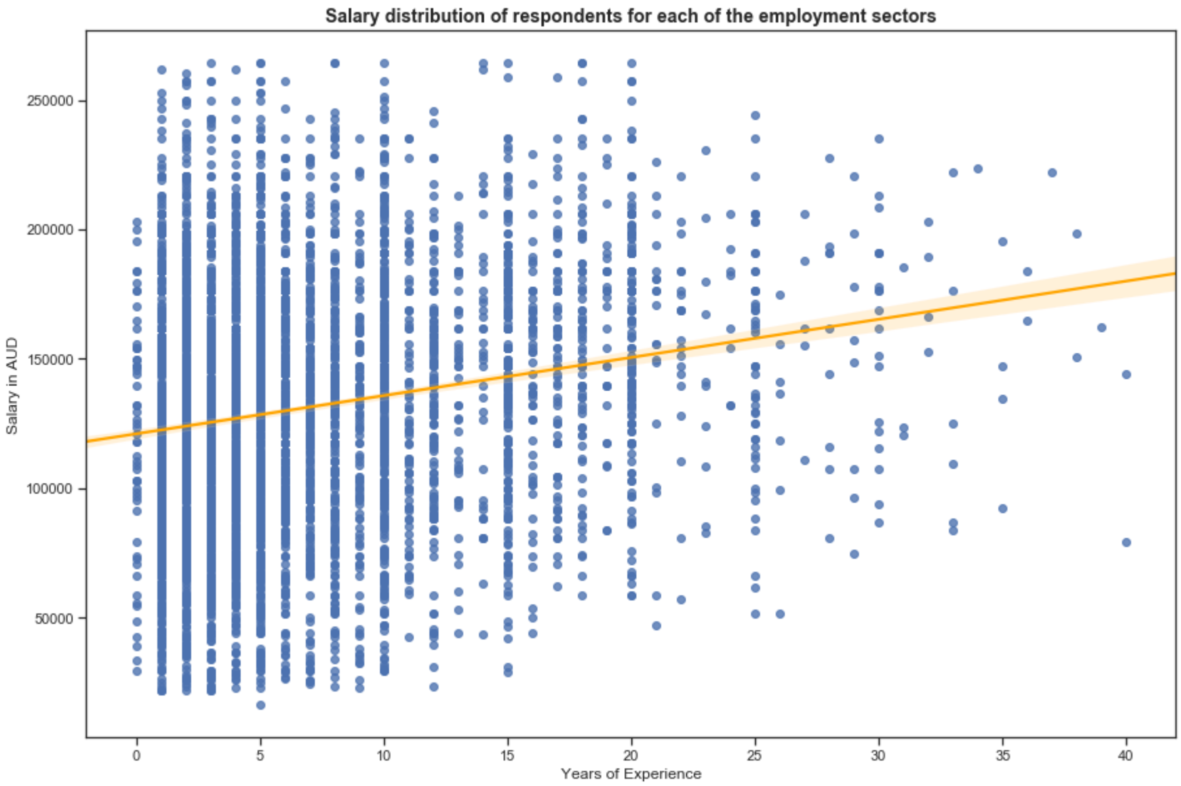
2. Let's refine this.
Remove Salary outliers using 2-sigma rule and then create a linear regression between the salary and years-of experience of full-time respondents.
Plot the linear fit over the scatterplot.

In [640]:

```
#Your code
#create the function to remove the outliers using 2-sigma rule using median
def remove_outliers(data,column, s=2):
     return data[abs(column - np.median(column)) < s * np.std(column)]
df = dataset_a1
column = df.SalaryAUD

#remove the outliers of Salary
data_no_outliers = remove_outliers(df,column)

plt.figure(figsize=(15,10)) #set the fig size
#plots the regplot for YearsofExperience and SalaryAUD
ax = sns.regplot(x="YearsofExperience",y="SalaryAUD",
data = data_no_outliers[data_no_outliers.EmploymentStatus ==
                        'Full time employee'],
line_kws = {'color':'orange'})
#set the title in bold and particular fontsize and the axis labels
ax.set_title(
"Salary distribution of respondents for each of the employment sectors",
             weight='bold').set_fontsize('14')
ax.set(ylabel = 'Salary in AUD' ,xlabel = 'Years of Experience')
```

Out[640]:

```
[Text(0, 0.5, 'Salary in AUD'), Text(0.5, 0, 'Years of Experience')]
```



@Benjamin Bannier. (May 13, 2013). Answer to question: Is there a numpy builtin to reject outliers from a list. Retrieved from: [https://stackoverflow.com/questions/11686720/is-there-a-numpy-builtin-to-reject-outliers-from-a-list (https://stackoverflow.com/questions/11686720/is-there-a-numpy-builtin-to-reject-outliers-from-a-list)](https://stackoverflow.com/questions/11686720/is-there-a-numpy-builtin-to-reject-outliers-from-a-list). Date accessed: Jan 18, 2020.

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

Answer Yes, I think this is a very good way of predicting salaries. As we are removing the outliers, it removes the random variation in the data and gives a focused dataset. Linear regression makes us easity to predict the dependency of salary on years of experience. The scatterplot helps to show the density(count) of respondents across the years of experience. Hence we can see that the salary increases as the years of experince increases. And, as the years of experience increaese, the count of respondents decreses.

6. Tasks and tools

You might be wondering (or not) what different tasks you will be assigned in a data science role and what kind of tools would you be using the most?

In this section, we perform necessary text processing to investigate such aspects.

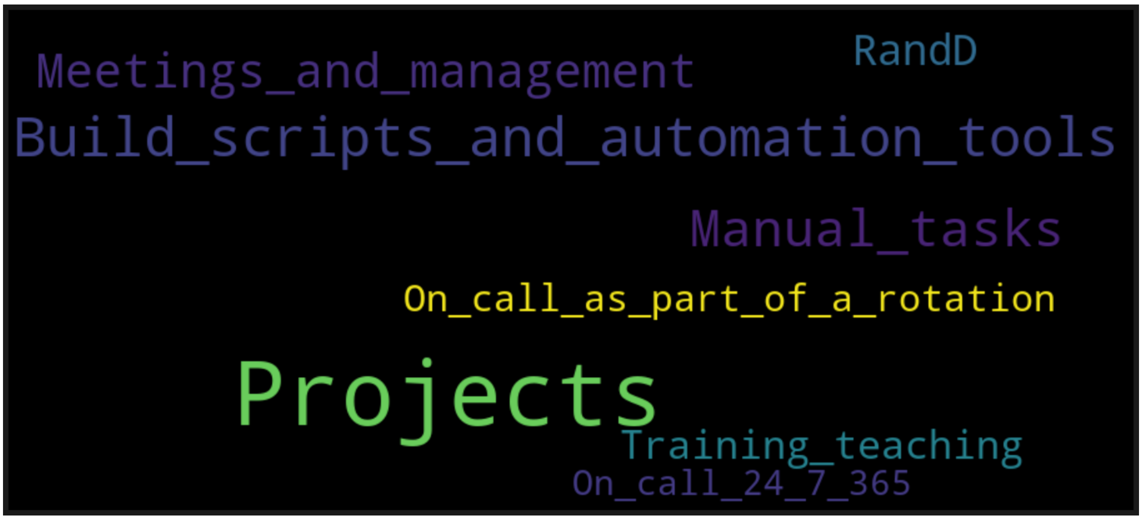
6.1 Data science common tasks

We focus here on the three data science job roles and investigate the tasks usually carried out in such roles.

1. Investigate the 'KindsOfTasksPerformed' column and perform the required text processing to enable you to plot a word cloud depicting the frequency of the different tasks.

In [599]:

```
# Your code

df=(dataset_a1['KindsOfTasksPerformed'])
#remove null values
df.replace(' ', np.nan, inplace = True)
df = df.dropna().copy(deep=True)
#convert column into string
dataset = df.astype(str)
#concatenate the string
string = ','.join(dataset)


#modify the string and remove the special
#chracters to get the desired pattern of words
final = string.replace(" ","_")
final = final.replace(",_"," ")
final = final.replace(","," ")
final = final.replace("&","and")
final = final.replace("/","_")
final_tasks = final.replace("-","_")
# final_tasks = final.replace("nan","")
dataset
#frequency of words
# def freq(str):

#     # break the string into list of words
#         str_list = final.split()

#     # gives set of unique words
#         unique_words = set(str_list)

#         count_list=[]

#         for words in unique_words :
#             count =  str_list.count(words)
#             count_list.append(count)
#         return  unique_words,count_list

#plot the wordcloud without repetitive words
wordcloud = WordCloud(
    width = 900,
    height = 400,
    background_color = 'black',
    collocations = False).generate(final_tasks)
fig = plt.figure(
    figsize = (20, 20),
    facecolor = 'k'
    )
plt.imshow(wordcloud, interpolation = 'bilinear')
plt.axis("off")
plt.tight_layout(pad=0)
plt.show()
```



@craigching. (Jul 14, 2017). Answer to question: Python: wordcloud, repetitve words. Retrieved from: https://stackoverflow.com/questions/43954114/python-wordcloud-repetitve-words (https://stackoverflow.com/questions/43954114/python-wordcloud-repetitve-words). Date accessed: Jan 20, 2020.

@PythonForBeginners. (Dec 09, 2012). Answer to question: String Concatenation and Formatting. Retrieved from: https://www.pythonforbeginners.com/concatenation/string-concatenation-and-formatting-in-python (https://www.pythonforbeginners.com/concatenation/string-concatenation-and-formatting-in-python). Date accessed: Jan 19, 2020.

@Bartosz Mikulski. (Aug 07, 2018). Answer to question: Word cloud from a Pandas data frame. Retrieved from:https://www.mikulskibartosz.name/word-cloud-from-a-pandas-data-frame/ (https://www.mikulskibartosz.name/word-cloud-from-a-pandas-data-frame/). Date accessed: Jan 19, 2020.

6.2 Data Science Common Tools

Now we compare the skillset required by data science roles and other IT roles.

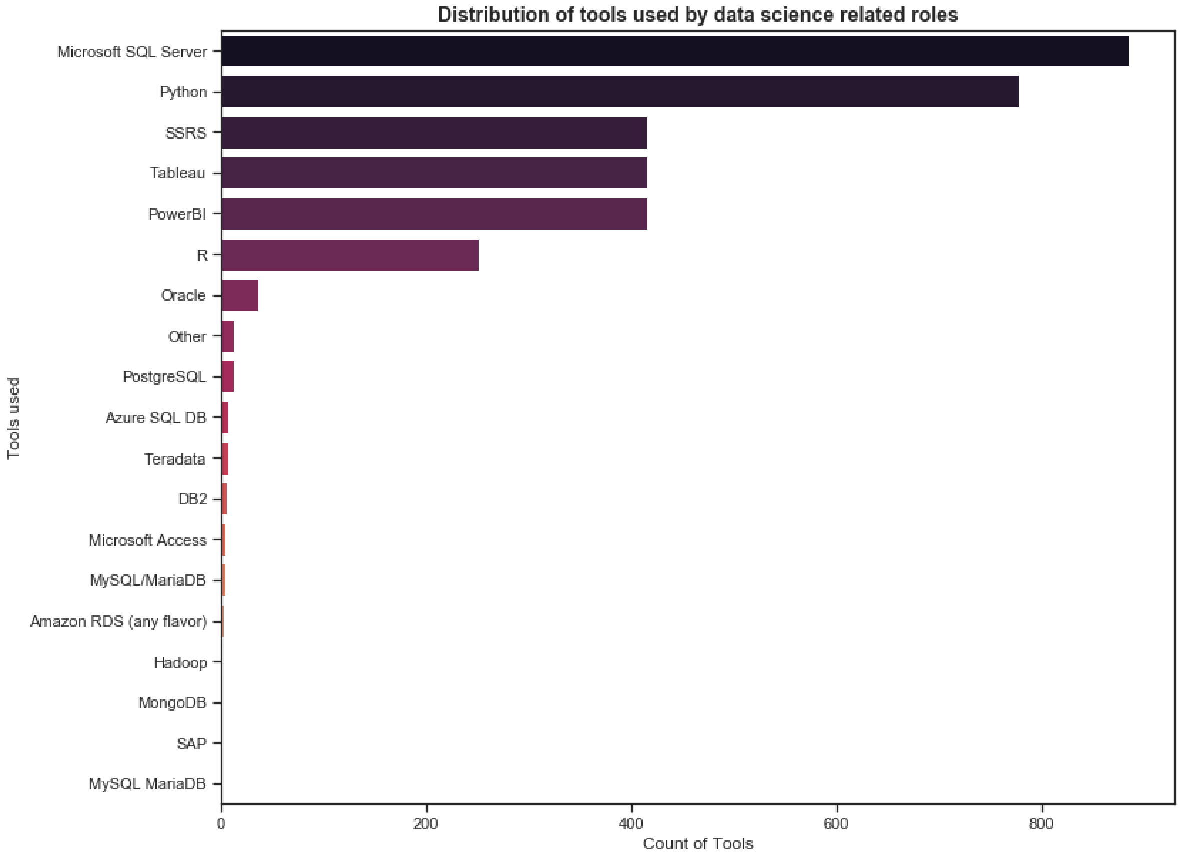
2. Filter your respondents based on DataScienceRelated flag and plot two seperate bar charts depicting the tools used by data science roles versus other roles.

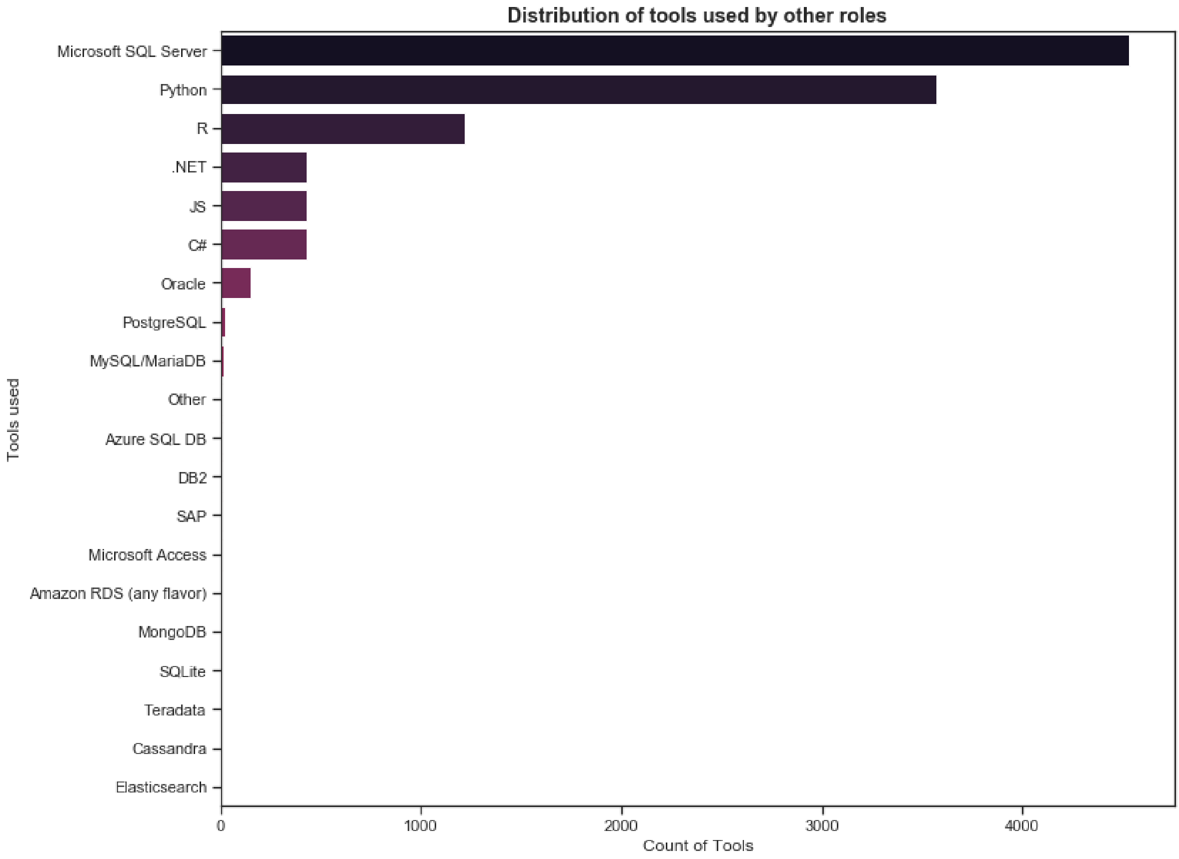
Hint: You will need to do similar text processing to the previous task.

In [639]:

```
# Your code
df = dataset_a1
#select data only for data science related roles
df_dsroles = df[df.DataScienceRelated == True]

#convert ToolsUsed column to list
dstools_list=(df_dsroles['ToolsUsed']).astype(str)

#split the list by delimiter ',' and internal arrays
dstools_list1 = list(map(lambda x: x.split(','), dstools_list))

#split the list internally by delimiter ',' and convert in string
dstools_string = ','.join(map(','.join, dstools_list1))

#split the string by delimiter ',' into single string objects
dstools_string1 = dstools_string.split(',')

#strip the white spaces and convert to map
dstools_final_string = map(str.strip,dstools_string1)
dstools_series= pd.Series(dstools_final_string) #convert into series

#select data only for other roles(non-data science related)
df_othroles = df[df.DataScienceRelated == False]

#convert ToolsUsed column to list
otools_list=(df_othroles['ToolsUsed']).tolist()

#split the list by delimiter ',' and internal arrays
otools_list1 = list(map(lambda x: x.split(','), otools_list))

#split the list internally by delimiter ',' and convert in string
otools_string = ','.join(map(','.join, otools_list1))

#split the string by delimiter ',' into single string objects
otools_string1 = otools_string.split(',')

#strip the white spaces and convert to map
otools_final_string = map(str.strip,otools_string1)
otools_series= pd.Series(otools_final_string) #convert into series

#plot the tools used by datasciencerelated job roles using countplot
plt.figure(figsize=(12,10)) #set the fig size

#plot the countplot for tools used by datascience related roles
ax=sns.countplot(y = dstools_series, palette = 'rocket',
             order = dstools_series.value_counts().index)
#set the title in bold and particular fontsize and the axis labels
ax.set(xlabel='Count of Tools', ylabel='Tools used')
plt.title('Distribution of tools used by data science related roles',
          weight='bold').set_fontsize('14')
plt.show(ax)

#plot the tools used by other job roles series using countplot
plt.figure(figsize=(12,10)) #set the fig size
#plots the countplot for tools used by other roles in order
bx =sns.countplot(y = otools_series, palette = 'rocket',
                  order = otools_series.value_counts().index)
#set the title in bold and particular fontsize and the axis labels
bx.set(xlabel='Count of Tools', ylabel='Tools used')
```

```
plt.title('Distribution of tools used by other roles',
          weight='bold').set_fontsize('14')
plt.show(bx)
```





@idjaw. (Oct 19, 2015). Answer to question: How do I convert multiple lists inside a list using Python? [duplicate]. Retrieved from:https://stackoverflow.com/questions/33223388/how-do-i-convert-multiple-lists-inside-a-list-using-python (https://stackoverflow.com/questions/33223388/how-do-i-convert-multiple-lists-inside-a-list-using-python). Date accessed: Jan 18, 2020.

3. What do you think are the most commonly used tools for a data science role?

Answer The most commonly used tools for data science role can be divided into three categories as follows: 1)Data Processing/Streaming Apache Hadoop Apache Spark Apache Pic Apache Hbase BigML (Data Handling) Cassandra Neo4j 2)Data Analysis: Excel R/Python 3)Data Visualization: Tableau D3 SAS Power BI tools

7. Data quality assessment

' Garbage in, garbage out'.

The saying means that poor quality data will return unreliable and often conflicting results. In this task, you need to assess your data set critically and understand not just what its use means for the outcome of your analysis, but also how those insights inform decisions which lead to broader effects.

1. Now that you have analysed the data. Go into the data set file and determine two anomalies. These could be parts of the data that don't seem quite right or logically can't co-exist. Write a paragraph about these explaining what part of your analysis alerted you to them, why they are anomalies, why they may exist, and what could be done to fix them.

Answer

Anomalies are the entries or observations which raise suspicions, exceptions and differ from the majority data. Two of the significant anomalies that brought out suspicious results are mentioned here. The first anomaly is the outlier which shows the highest salary of 96000 USD, is just an associate who does not even have education computer related. The second highest salary respondent shows similar unusual behaviour with age 23 and no degree completed, having experience of just 5 years. Hence these cause suspicion and unusual behaviours than the results occupied when we remove the outliers for relating the years of experience, age, education and salary statistics. This anomaly was discovered during analysing the data and results for plotting the formal education type vs the salary distribution boxplot which gets plotted as outliers in the graph. This is in contrast with the analysis that one should have a significant years of experience,or atleast computer related formal degree(Bachelors or Masters) to get paid well. The cause of such anomaly may be just false or fake data input. The data is not verified with any formal proof etc., hence can be easily faked and is hard to be trusted. The second anomaly is the salary distribution in Uganda for the 4 entries which shows unrealistic salaries in the range from 71000 to 185000 USD i.e. approx 679 million UGX(currency of Uganda). This is hard to believe for a country who is not so economically developed. Also as verified from google, this data is significantly high to the highes salary stated in Uganda. This has also created misleading results while plotting the graph of salaries according to the countries. The question which asks to analyse the graph for countries with the highest salary alerted me for this anomaly. The reason for this anomaly might be that the Salary field is in USD, but in this case might have been entered in the wrong currency. This anomalies can be fixed by detecting them first by implementing the most suitable anomaly detection technique and then removing them or changing/ adjusting the values to the something that is more representative of our dataset. They can also add constraints to the fields or modify the method the data is collected by adding some verification or proffing methods that will help to gain accurate and genuine data.

Well done! You have completed the assignment!

For reassurance, the Australian 2019 Graduate Outcomes Survey found the median salary for Masters graduates in Computer Science and Information Systems for was AUD 92,900 for full-time employment.