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: Zexi Liu

Student number: 29295181

Tutorial number. :10-P1

Tutor: Abdullah alghamdi

Environment: Python *(3.7.4)* Anaconda 1.9.7 *(64-bit)*

In [4]:

```
from platform import python_version

print(python_version())
```

```
3.7.4
```

Bastien,L. (2009). Printing Python version in output. Retrieved from https://stackoverflow.com/questions/1252163/printing-python-version-in-output (https://stackoverflow.com/questions/1252163/printing-python-version-in-output)

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 [5]:

```
# Your code
!pip install seaborn
#install visualization library
!pip install pandas
#install data frame analysis library
import seaborn as sns
#import visualization library
import pandas as pd
#import data frame analysis library

```

```
Requirement already satisfied: numpy>=1.9.3 in /opt/anaconda3/lib/pyt
hon3.7/site-packages (from seaborn) (1.17.2)
Requirement already satisfied: matplotlib>=1.4.3 in /opt/anaconda3/li
b/python3.7/site-packages (from seaborn) (3.1.1)
Requirement already satisfied: pytz>=2017.2 in /opt/anaconda3/lib/pyt
hon3.7/site-packages (from pandas>=0.15.2->seaborn) (2019.3)
Requirement already satisfied: python-dateutil>=2.6.1 in /opt/anacond
a3/lib/python3.7/site-packages (from pandas>=0.15.2->seaborn) (2.8.0)
Requirement already satisfied: cycler>=0.10 in /opt/anaconda3/lib/pyt
hon3.7/site-packages (from matplotlib>=1.4.3->seaborn) (0.10.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /opt/anaconda3/li
b/python3.7/site-packages (from matplotlib>=1.4.3->seaborn) (1.1.0)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.
0.1 in /opt/anaconda3/lib/python3.7/site-packages (from matplotlib>=
1.4.3->seaborn) (2.4.2)
Requirement already satisfied: six>=1.5 in /opt/anaconda3/lib/python
3.7/site-packages (from python-dateutil>=2.6.1->pandas>=0.15.2->seabo
rn) (1.12.0)
Requirement already satisfied: setuptools in /opt/anaconda3/lib/pytho
n3.7/site-packages (from kiwisolver>=1.0.1->matplotlib>=1.4.3->seabor
```

Aditya,K. (2009). Python & pip Windows installation. Retrieved from https://github.com/BurntSushi/nfldb/wiki/Python-&-pip-Windows-installation (https://github.com/BurntSushi/nfldb/wiki/Python-&-pip-Windows-installation)

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.

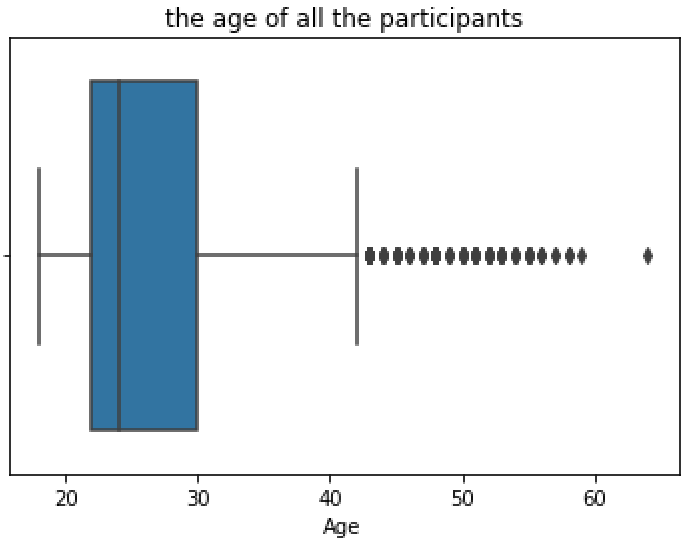
1. Create a box plot showing the age of all the participants.
Your plot must have labels for each axis, a title, numerical points for the age axis and also show the outliers.

In [6]:

```
# Your code
df = pd.read_csv('assignment1_dataset.csv',sep=',') #import data from csv files
sns.boxplot(df['Age']).set_title('the age of all the participants') #draw boxpl
```

Out[6]:

```
Text(0.5, 1.0, 'the age of all the participants')
```



Shanelynn. (2019). Python Pandas read_csv – Load Data from CSV Files. Retrieved from [https://www.shanelynn.ie/python-pandas-read_csv-load-data-from-csv-files/ (https://www.shanelynn.ie/python-pandas-read_csv-load-data-from-csv-files/)](https://www.shanelynn.ie/python-pandas-read_csv-load-data-from-csv-files/)

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 [7]:

```
# Your code
df.Age.quantile([0,0.25,0.5,0.75,1]) #quatile 0.5 is median not mean
```

Out[7]:

```
0.00    18.0
0.25    22.0
0.50    24.0
0.75    30.0
1.00    64.0
Name: Age, dtype: float64
```

Shubham,R(2019).Python | Pandas dataframe.quantile().Retrieved from [https://www.geeksforgeeks.org/python-pandas-dataframe-quantile/ (https://www.geeksforgeeks.org/python-pandas-dataframe-quantile/)](https://www.geeksforgeeks.org/python-pandas-dataframe-quantile/)

In [8]:

```
round(df.Age.mean())#round age mean
```

Out[8]:

```
27
```

Answer

minimum:18

first quartile:22

median:27

third quartile:30

maximum:64

mean:27

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 From the whisker we can firstly see the max age and the min age of the participants, which are 18 years old for the min and 64 years old for the max (but is not the Q4 here), thus, the range of the age is 46 years old. From the median which is 27 years old, we can know that half of participants are great than 27 and other half are younger than 27. And the first quartile means the middle of all of the ages younger than 27, and the third quartile means the middle of all of the age older than 27. And the interquartile range(IQR) is third quartile minus first quartile. In this case is 8. To calculate the outliers, we must first calculate the Q1-1.5*IQR which is 10 in this stage. Because 18 is greater than 10, it is accepted. Using the similar method, Q3+1.5IQR it's 42. Thus, the age from 42 to 64 are all outliers.

Answer ii Here, from the boxplot we can see Q3-Q2 > Q2-Q1 and the whisker on the right is longer than the left whisker, thus, the data is skewed right. From this condition, the mode is less than the mean.

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 [9]:

```
# Your code
df.Age.loc[(df.Age > 65)|(df.Age <20)].count()#locate the data which is great t
```

Out[9]:

```
90
```

Shubham,R.(2019). Python | Pandas DataFrame.loc[] Retrieved from [https://www.geeksforgeeks.org/python-pandas-dataframe-loc/ (https://www.geeksforgeeks.org/python-pandas-dataframe-loc/)](https://www.geeksforgeeks.org/python-pandas-dataframe-loc/)

Answer
The number of respondents is 90.

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.

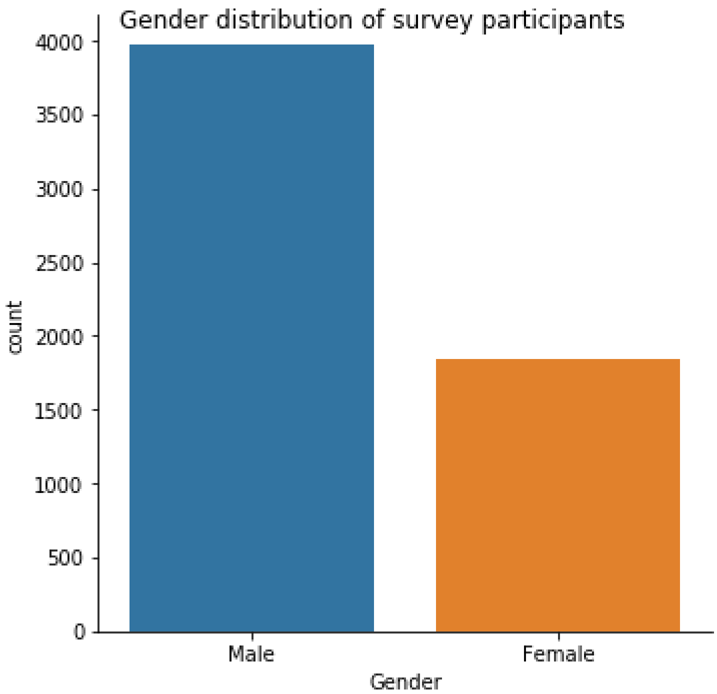
5. Plot the gender distribution of survey participants.

In [10]:

```
# Your code
g = sns.catplot(x = 'Gender',kind = 'count',data = df,margin_titles = True) #co
g.fig.suptitle('Gender distribution of survey participants') #add title
```

Out[10]:

```
Text(0.5, 0.98, 'Gender distribution of survey participants')
```



Mwaskom (2020). seaborn FacetGrid: How to leave proper space on top for suptitle Retrieved from [https://stackoverflow.com/questions/28638158/seaborn-facetgrid-how-to-leave-proper-space-on-top-for-suptitle (https://stackoverflow.com/questions/28638158/seaborn-facetgrid-how-to-leave-proper-space-on-top-for-suptitle)](https://stackoverflow.com/questions/28638158/seaborn-facetgrid-how-to-leave-proper-space-on-top-for-suptitle)

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

In [11]:

```
# Your code
df.Gender.value_counts(normalize = True)#relative frequencies of the unique val
```

Out[11]:

```
Male      0.683523
Female    0.316477
Name: Gender, dtype: float64
```

Parul,P (2019). Getting more value from the Pandas' value_counts() retrieved from [https://towardsdatascience.com/getting-more-value-from-the-pandas-value-counts-aa17230907a6 (https://towardsdatascience.com/getting-more-value-from-the-pandas-value-counts-aa17230907a6)](https://towardsdatascience.com/getting-more-value-from-the-pandas-value-counts-aa17230907a6)

Answer
the man's percentage is 68.3523% and the woman's percentage is 31.6477%

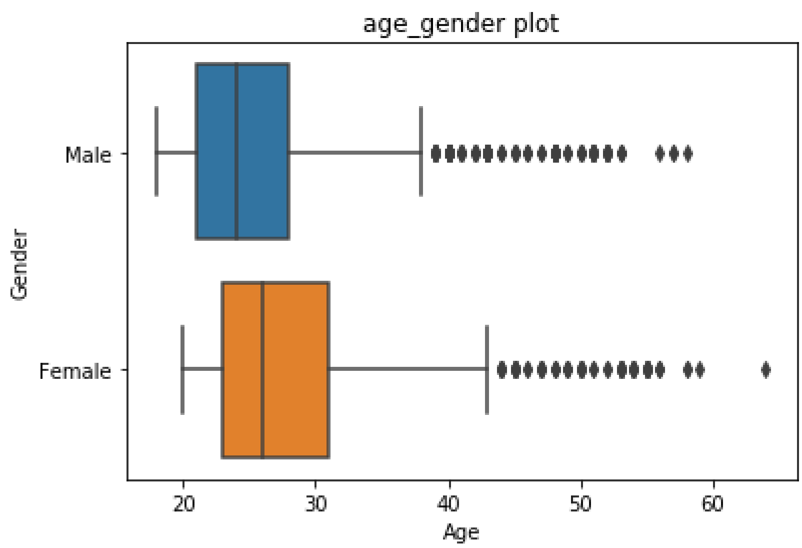
7. Let's see if there is any relationship between age and gender.
Create a box plot showing the age of all the participants according to gender.

In [22]:

```
# Your code
sns.boxplot(x = 'Age', y = 'Gender',data = df).set_title('age_gender plot')#age
```

Out[22]:

```
Text(0.5, 1.0, 'age_gender plot')
```



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 [23]:

```
# Your code
Male = df.loc[df.Gender == 'Male']
Male = Male.describe()
Male = Male.rename(columns={"Age" : "descriptive_statistics"})
Male.descriptive_statistics #only display what customer expected
```

Out[23]:

```
count    3974.000000
mean       26.058128
std         6.513043
min        18.000000
25%        21.000000
50%        24.000000
75%        28.000000
max        58.000000
Name: descriptive_statistics, dtype: float64
```

pandas: Rename index / columns names (labels) of DataFrame(2019,February) in nkmk. Retrieved from https://note.nkmk.me/en/python-pandas-dataframe-rename/ (https://note.nkmk.me/en/python-pandas-dataframe-rename/)

In [24]:

```
Female = df.loc[df.Gender == 'Female']
Female = Female.describe()
Female = Female.rename(columns = {"Age" : "descriptive_statistics"})
Female.descriptive_statistics
```

Out[24]:

```
count    1840.000000
mean       28.285870
std         7.646899
min        20.000000
25%        23.000000
50%        26.000000
75%        31.000000
max        64.000000
Name: descriptive_statistics, dtype: float64
```

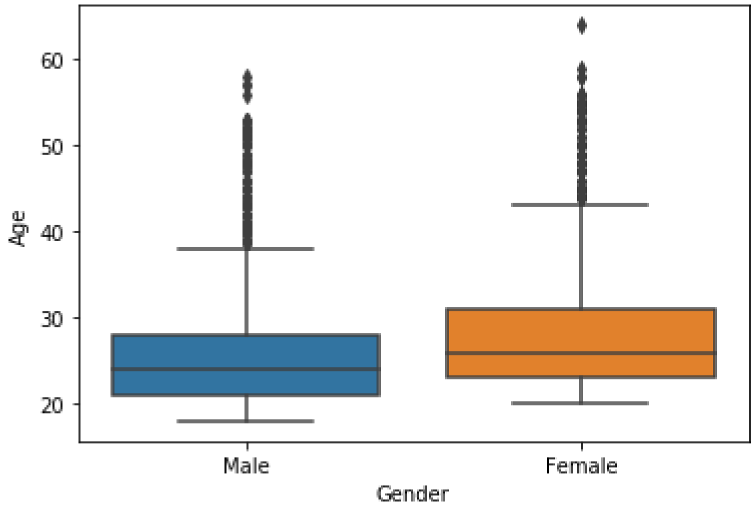
Varun(2018,November). Using pandas describe method to get dataframe summary Retrieved from https://backtobazics.com/python/pandas-describe-method-dataframe-summary/ (https://backtobazics.com/python/pandas-describe-method-dataframe-summary/)

In [25]:

```
sns.boxplot(x = 'Gender',y = 'Age',data = df)
```

Out[25]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a1b68bd90>
```



Answer

male's descriptive statistics shape is similar to female one, but number of five descriptive statics for male are lower than female. The attribute of male are mean=26,std=6.513043,min=18,25%=21,50%=24,75%=28,max=58 The attribute of female are mean=28,std=7.646899,min=20,25%=23,50%=26,75%=31,max=64

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 [154]:

```
# Your code
top = df.Country.value_counts(normalize = True)[:30]#find percentage, default s
Top30 = sns.barplot(top.index,top)#corresponding the argument one by one
Top30.set_xticklabels(Top30.get_xticklabels(), rotation=80)#rotate x label 80 d
```

```
 Text(0, 0, 'Czech Republic'),
 Text(0, 0, 'Austria'),
 Text(0, 0, 'Hungary'),
 Text(0, 0, 'Turkey'),
 Text(0, 0, 'Portugal'),
 Text(0, 0, 'Mexico')]
```

ImportanceOfBeingErnest (2017). Rotate xtick labels in seaborn boxplot? Retrieved from https://stackoverflow.com/questions/44954123/rotate-xtick-labels-in-seaborn-boxplot (https://stackoverflow.com/questions/44954123/rotate-xtick-labels-in-seaborn-boxplot)

In [28]:

```
print (top.round(4)*100)#calculate the percentage and round two decimal
```

```
United States     66.15
United Kingdom    10.04
Canada             3.61
Australia          2.87
Sweden             1.32
Germany            1.29
Netherlands        1.15
India              1.12
South Africa       0.81
New Zealand        0.79
Denmark            0.71
Poland             0.67
Switzerland        0.60
Romania            0.57
Ireland            0.55
France             0.55
Spain              0.52
Italy              0.52
Russia             0.48
Belgium            0.43
Norway             0.36
Israel             0.34
Brazil             0.34
Finland            0.29
Czech Republic     0.28
Austria            0.26
Hungary            0.21
Turkey             0.21
Portugal           0.21
Mexico             0.21
Name: Country, dtype: float64
```

Answer The top five country are United States 66.15% United Kingdom 10.04% Canada 3.61% Australia 2.87% Sweden 1.32%

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 [29]:

```
# Your code
TopFiveCount = round(df.Country.value_counts(normalize = True).sort_values(ascen
TopFiveCount
```

Out[29]:

```
United States     66.15
United Kingdom    10.04
Canada             3.61
Australia          2.87
Sweden             1.32
Name: Country, dtype: float64
```

Answer The top five country are United States 66.15% United Kingdom 10.04% Canada 3.61% Australia 2.87% Sweden 1.32%

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 United States holds more than half of the percentage of the job opportunity, it's 66.15%, United Kingdom follows USA, have 10% which place the second position. Germany ranks the sixth position which is 1.29%. The place of USA and UK is as expected, however, the need of Germany is less than Canada, Australia and Sweden is unexpected.

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 [30]:

```
TopFiveCountDf = df[(df.Country == 'United States')|(df.Country == 'United Kingd
TopFiveCountDf = TopFiveCountDf.groupby(['Country','Gender'])['Age']#group coun
```

Brad,S (2019). Pandas GroupBy: Your Guide to Grouping Data in Python Retrieved from [https://realpython.com/pandas-groupby/ (https://realpython.com/pandas-groupby/)](https://realpython.com/pandas-groupby/)

In [31]:

```
TopFiveCountDf.mean()
```

Out[31]:

```
Country         Gender
Australia       Female    27.863636
                Male      26.902439
Canada          Female    26.658537
                Male      26.869822
Sweden          Female    27.050000
                Male      26.912281
United Kingdom  Female    25.963415
                Male      24.754762
United States   Female    28.709538
                Male      26.310317
Name: Age, dtype: float64
```

In [32]:

```
TopFiveCountDf.median()
```

Out[32]:

```
Country         Gender
Australia       Female    26.5
                Male      25.0
Canada          Female    25.0
                Male      25.0
Sweden          Female    25.0
                Male      25.0
United Kingdom  Female    24.0
                Male      22.0
United States   Female    26.0
                Male      23.5
Name: Age, dtype: float64
```

In [33]:

```
TopFiveCountDf.count()
```

Out[33]:

```
Country         Gender
Australia       Female      44
                Male       123
Canada          Female      41
                Male       169
Sweden          Female      20
                Male        57
United Kingdom  Female     164
                Male       420
United States   Female    1384
                Male      2462
Name: Age, dtype: int64
```

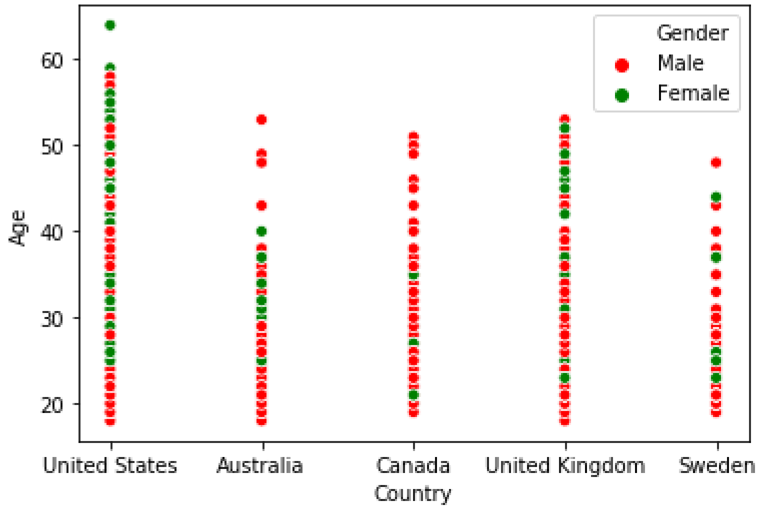
13. What Pattern do you notice about the relationship between age, gender for each of these countries? (if any).

In [34]:

```
TopFiveCountDf = df[(df.Country == 'United States')|(df.Country == 'United King
sns.scatterplot(x="Country", y="Age",hue="Gender", palette=["r", "g"],data=TopF
```

Out[34]:

<matplotlib.axes._subplots.AxesSubplot at 0x1a1c180f10>



seaborn.scatterplot(2020). seaborn 0.9.0 retrieved from https://seaborn.pydata.org/generated/seaborn.scatterplot.html (https://seaborn.pydata.org/generated/seaborn.scatterplot.html)

Answer From the image, we can see that USA Female and Male are balanced at every age range. All age goes from 20 to 60. Australia's female are most likely between 30 and 40 range. The number of male data over 40 become decreased. Canada do not have much female, which located at 20 to 40 range. UK's female range are mostly located in range of 40 to 50. Last but not least, sweden's female located at 20 to 30, and their male over 30 years old becomes decreased.

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.

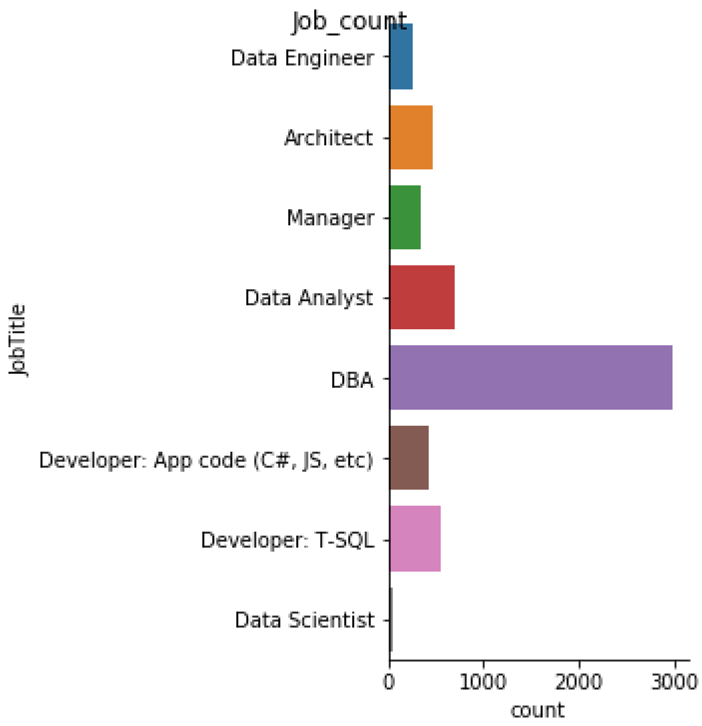
14. Plot a bar graph depicting the counts of different roles (each bar should represent the count of participants assuming a certain job role).

In [35]:

```
# Your code
roles = sns.catplot(y = 'JobTitle',kind = 'count',data = df);
roles.fig.suptitle('Job_count') #add title
```

Out[35]:

```
Text(0.5, 0.98, 'Job_count')
```



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

In [36]:

```
# Your code
(df.JobTitle.value_counts(normalize = True))*100
```

Out[36]:

```
DBA                                  51.547988
Data Analyst                         12.091503
Developer: T-SQL                      9.666323
Architect                             8.015136
Developer: App code (C#, JS, etc)     7.327141
Manager                               5.882353
Data Engineer                         4.643963
Data Scientist                        0.825593
Name: JobTitle, dtype: float64
```

Answer the percentage of the data scientist is 0.8256%

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 [16]:

```
# Your code
DataScienceRelated = (df.JobTitle == 'Data Engineer')|(df.JobTitle == 'Data Scie
DataScienceRelated
```

Out[16]:

```
0        True
1       False
2       False
3       False
4        True
        ...
5809     True
5810     True
5811     True
5812     True
5813     True
Name: JobTitle, Length: 5814, dtype: bool
```

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

In [19]:

```
# Your code
df['DataScienceRelated']=DataScienceRelated
(df.DataScienceRelated.value_counts(normalize = True))*100

```

Out[19]:

```
0        True
1       False
2       False
3       False
4        True
        ...
5809     True
5810     True
5811     True
5812     True
5813     True
Name: JobTitle, Length: 5814, dtype: bool
```

Answer

the percentage of Data Science related roles among the survey participants is 17.5611%

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

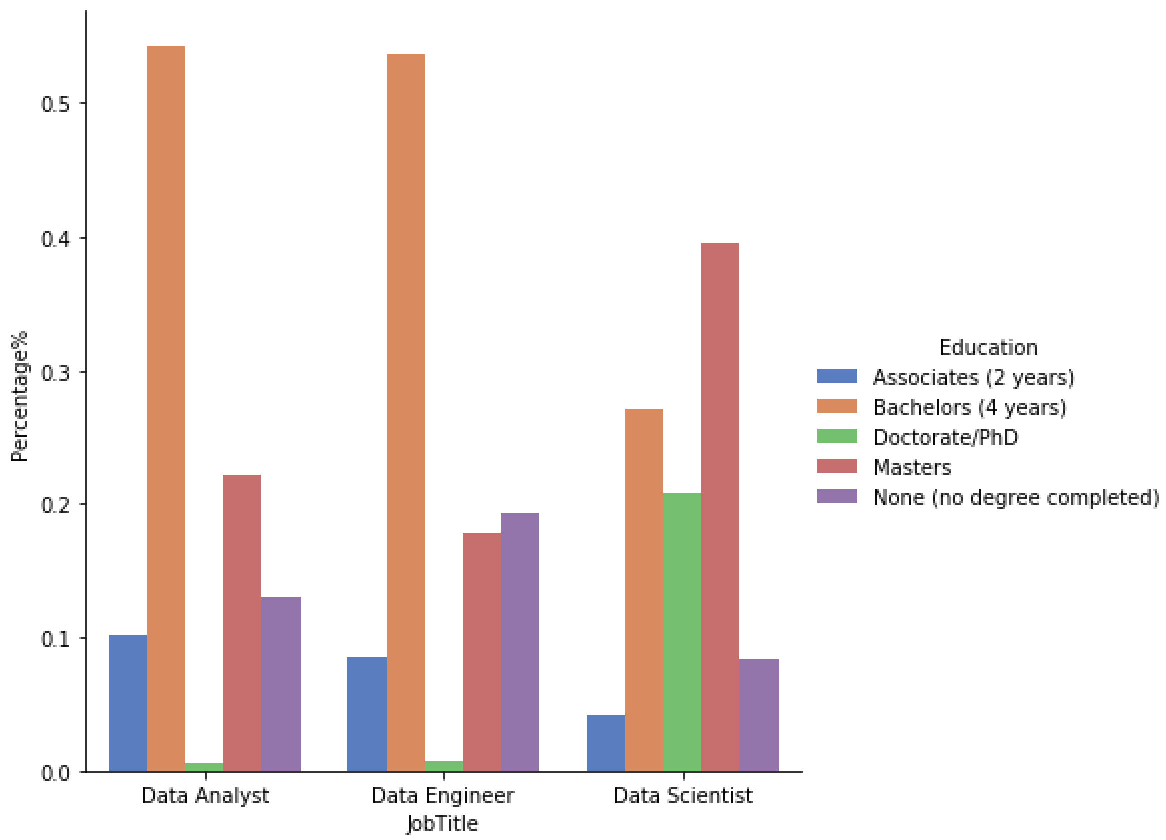
We saw in a recent activity that a significant number of data scientists job advertisements call for a masters degree or a PhD. Let's see if this is a reasonable ask based on the respondent's formal education.

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 [21]:

```
#your code
groupSeries = df[DataScienceRelated].groupby(['JobTitle','Education']).count().:
col_val = {}
col_val[groupSeries.index.get_level_values(0).name] = groupSeries.index.get_leve
col_val[groupSeries.index.get_level_values(1).name] = groupSeries.index.get_leve
col_val['Count'] = groupSeries.values
groupDf = pd.DataFrame(col_val)
total = groupDf.groupby('JobTitle').sum().Count.to_dict()
groupDf['Total'] = pd.Series([total[t] for t in groupDf.JobTitle])
groupDf['Percentage%'] = groupDf['Count']/groupDf['Total']
g = sns.catplot(x="JobTitle", y="Percentage%", hue="Education", data=groupDf,he:
```



Paul, H (2020). pandas - multi index plotting Retrieved from
https://stackoverflow.com/questions/31845258/pandas-multi-index-plotting

https://stackoverflow.com/questions/31845258/pandas-multi-index-plotting
(https://stackoverflow.com/questions/31845258/pandas-multi-index-plotting)

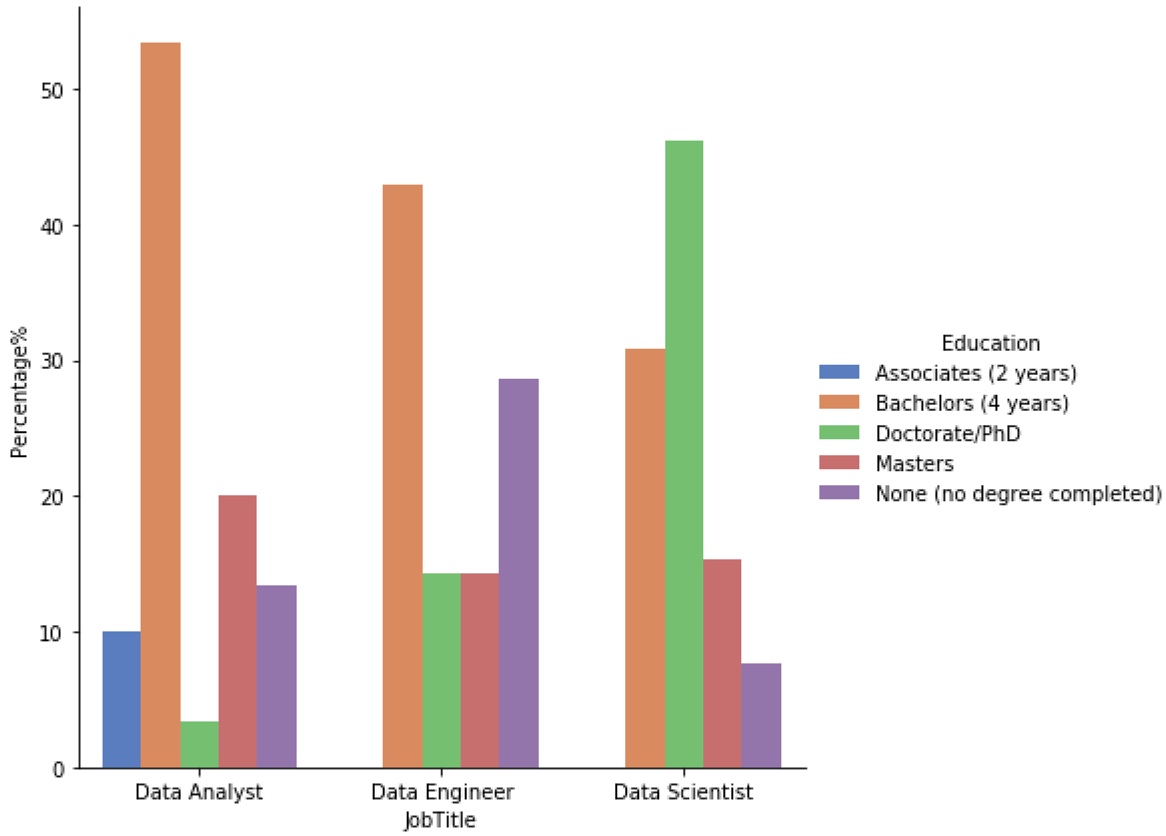
2. Based on what you have seen, do you think that a Master's or Doctoral degree is too unrealistic for job advertisers looking for someone with data science skills or is it job-dependent?

Answer from the image and chart, we can see that data science need 20% of PhD 40% of Masters. Data engineer need 18%of masters and 0.7%of PhD. Data analyst need 22% of masters and 0.5% of PhD. The need will be increased overtime if this industry is still booming.

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 [30]:

```
AustraliaDataScienceRelated = ((df.JobTitle == 'Data Engineer')|(df.JobTitle ==
dataRelated = df[AustraliaDataScienceRelated]
groupSeries = dataRelated.groupby(['JobTitle','Education']).count().iloc[:,0]
col_val = {}
col_val[groupSeries.index.get_level_values(0).name] = groupSeries.index.get_leve
col_val[groupSeries.index.get_level_values(1).name] = groupSeries.index.get_leve
col_val['Count'] = groupSeries.values
groupDf = pd.DataFrame(col_val)
total = groupDf.groupby('JobTitle').sum().Count.to_dict()
groupDf['Total'] = pd.Series([total[t] for t in groupDf.JobTitle])
groupDf['Percentage%'] = groupDf['Count']/groupDf['Total']*100
g = sns.catplot(x="JobTitle", y="Percentage%", hue="Education", data=groupDf,he:
```



Answer Count DBA is 3m Data Analyst is 4, Data Engineer is 2, Data Sciencetist is 10.

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

In [43]:

```
# Your code
EducationGroup = df.groupby(['Education'])
print(EducationGroup[['Education','Age']].mean())
```

```
                                Age
Education
Associates (2 years)      26.447077
Bachelors (4 years)       26.682612
Doctorate/PhD             31.363636
Masters                   27.524449
None (no degree completed)  26.158746
```

In [44]:

```
EducationGroup = df.groupby(['Education'])
print(EducationGroup[['Education','Age']].median())
```

```
                            Age
Education
Associates (2 years)         24
Bachelors (4 years)          24
Doctorate/PhD                29
Masters                      25
None (no degree completed)   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.

In [42]:

```
# Your code
AustraliaDataScienceRelated = df[(df.JobTitle == 'Data Engineer')|(df.JobTitle =
DoctoralDegree = AustraliaDataScienceRelated[AustraliaDataScienceRelated.Educati
DoctoralDegree = DoctoralDegree.groupby(['JobTitle','Education']).count()
DoctoralDegree = DoctoralDegree.rename(columns={"Age" : "Count"})
DoctoralDegree.Count
```

Out[42]:

```
JobTitle        Education
DBA             Doctorate/PhD     3
Data Analyst    Doctorate/PhD     4
Data Engineer   Doctorate/PhD     2
Data Scientist  Doctorate/PhD    10
Name: Count, dtype: int64
```

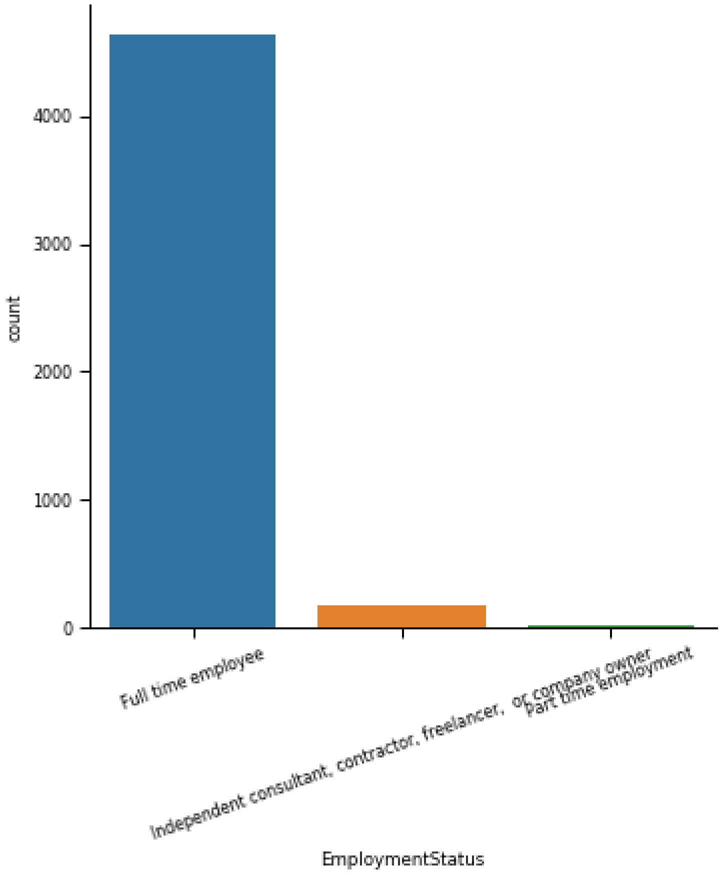
1. Plot the type of employment the respondents have on a bar chart for respondents who do not assume data science related roles.

In [82]:

```
# Your code
NoDataScience = df[(df.JobTitle != 'Data Engineer')&(df.JobTitle != 'Data Analys
NoDataScienceEmployment = sns.catplot(x = 'EmploymentStatus',kind = 'count',data
NoDataScienceEmployment.set_xticklabels(rotation=20)
```

Out[82]:

<seaborn.axisgrid.FacetGrid at 0x1a1e253b10>



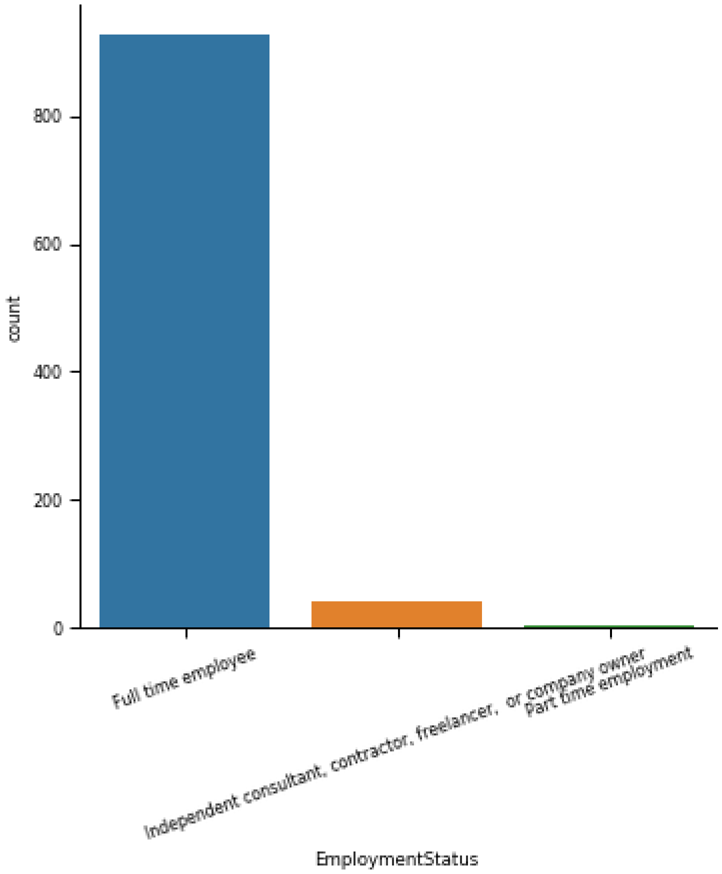
2. Now plot the type of employment the respondents have on a bar chart only for those assuming data science related roles

In [83]:

```
# Your code
DataScience = df[(df.JobTitle == 'Data Engineer')|(df.JobTitle == 'Data Analyst
#EmploymentStatus
sns.set_context("paper",font_scale = 0.9)
DataScienceEmployment = sns.catplot(x = 'EmploymentStatus',kind = 'count',data =
DataScienceEmployment.set_xticklabels(rotation=20)
```

Out[83]:

```
<seaborn.axisgrid.FacetGrid at 0x1a1e344650>
```



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 From the image, we can see, two figures have similar distribution. Thus, we cannot say data science roles differ in the type of employment as opposed to non-data science roles.

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 [47]:

```
# Your code
employmentCountry = df[(df.Country == 'Australia')|(df.Country == 'United Kingd
employmentCountry.Country.value_counts(normalize = True)*100
```

Out[47]:

```
United States     83.333333
United Kingdom    12.960497
Australia          3.706170
Name: Country, dtype: float64
```

Answer the percentages of full time employment in Australia is 3.7% in United Kingdom is 12.96% in United States is 83.33%

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.

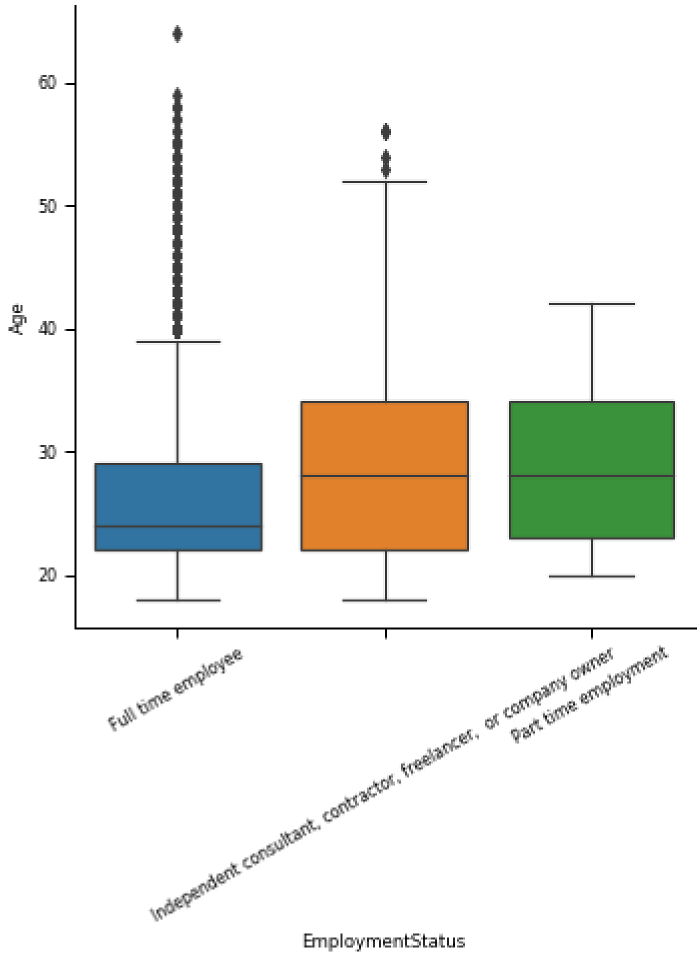
5. Plot a boxplot of the respondents age, grouped by employment type.

In [48]:

```
# Your code
ageEmployment = sns.catplot (y = 'Age', x = 'EmploymentStatus',kind = 'box',data
ageEmployment.set_xticklabels(rotation=30)
```

Out[48]:

<seaborn.axisgrid.FacetGrid at 0x1a1bbbb9d0>



6. What are your observations?

Answer there are too many outliers of full time employee over 40 years old. For independent or company owner, only a little bit outlier. And partime employment distributed evenly.

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".

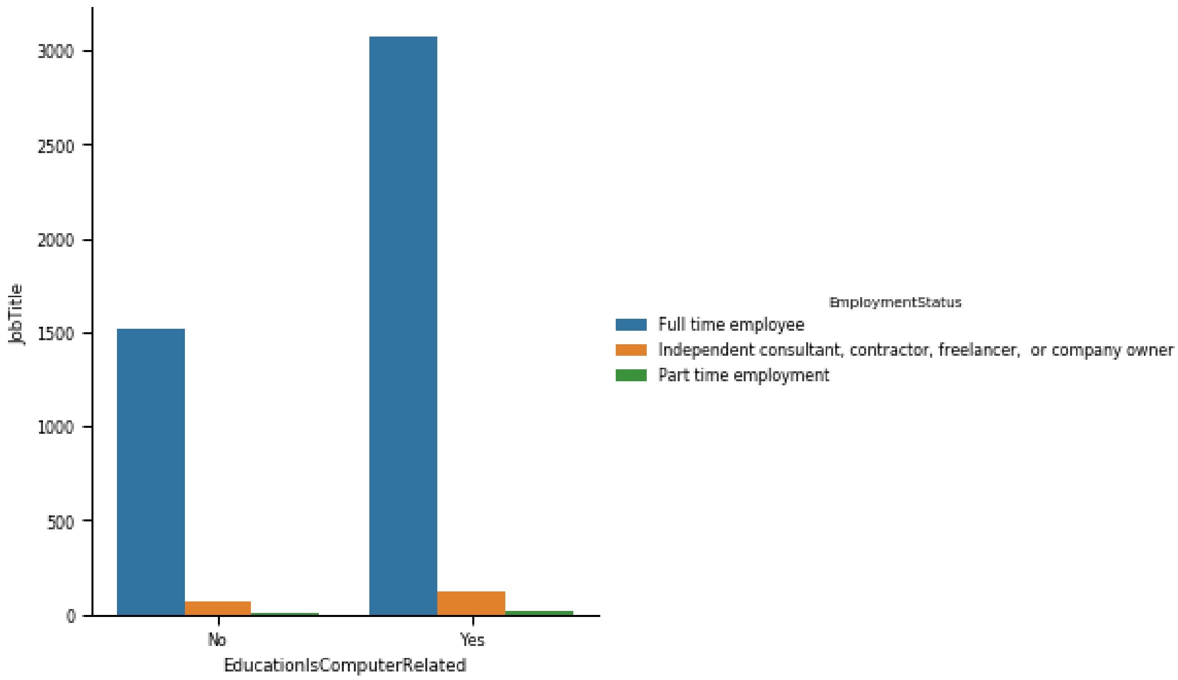
In [49]:

```
# Your code
computerRelatedStatus = df.groupby(['EducationIsComputerRelated','EmploymentStat
computerRelatedStatus = computerRelatedStatus.reset_index(level=0)
computerRelatedStatus = computerRelatedStatus.reset_index(level=0)
sns.catplot(x = 'EducationIsComputerRelated',y = 'JobTitle',hue = 'EmploymentSta

```

Out[49]:

```
<seaborn.axisgrid.FacetGrid at 0x1a1c3dda10>
```



pandas.Series.reset_index(2020). pandas 0.25.3 documentation Retrieved from [https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.reset_index.html (https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.reset_index.html)](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.reset_index.html)

In [50]:

```
computerRelatedStatus = df.groupby(['EducationIsComputerRelated','EmploymentStat
computerRelatedStatus = computerRelatedStatus.rename(columns={"Age" : "Count"})
computerRelatedStatus.Count
```

Out[50]:

```
EducationIsComputerRelated  EmploymentStatus
No                          Full time employee
1522
                            Independent consultant, contractor, freela
ncer,  or company owner      66
                            Part time employment
7
Yes                         Full time employee
3071
                            Independent consultant, contractor, freela
ncer,  or company owner     117
                            Part time employment
12
Name: Count, dtype: int64
```

8. Looking at the graph, does holding a computer-related degree improves your chances of securing a full-time job?
Explain your answers.

Answer Yes. From the figure,we can find that almost all people with computer degree background are more likely to find full time jobs.

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.

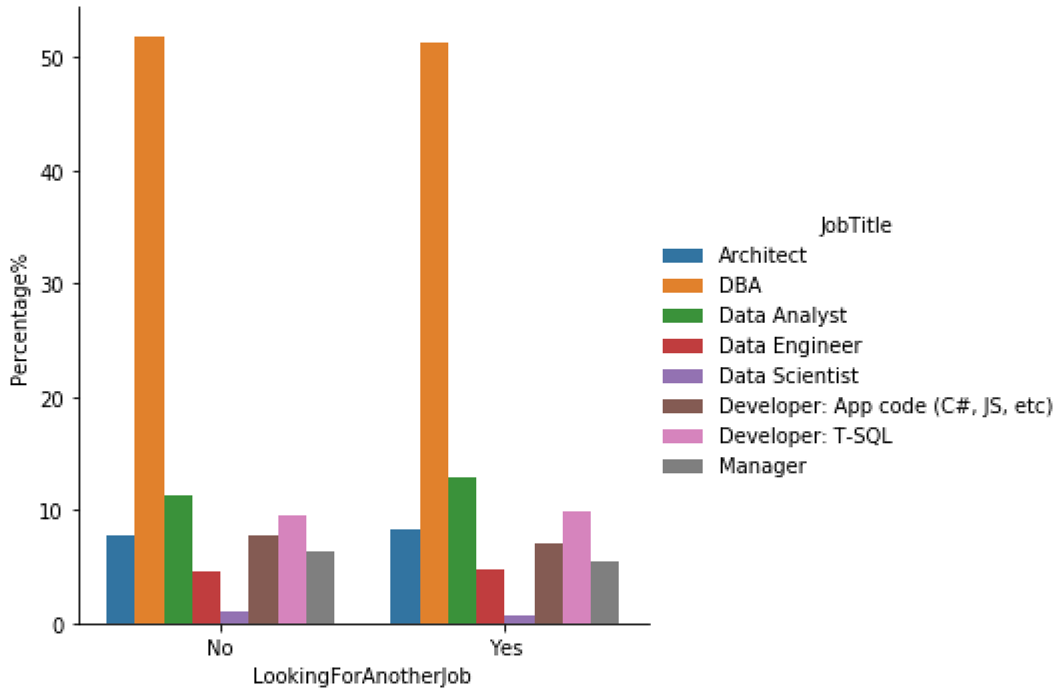
9. Create a bar chart for the percentage of respondents who are looking for another job grouped by the different job titles.

In [29]:

```
# Your code
groupSeries = df.groupby(['JobTitle','LookingForAnotherJob']).count().iloc[:,0]
col_val = {}
col_val[groupSeries.index.get_level_values(0).name] = groupSeries.index.get_leve
col_val[groupSeries.index.get_level_values(1).name] = groupSeries.index.get_leve
col_val['Count'] = groupSeries.values
groupDf = pd.DataFrame(col_val)
total = groupDf.groupby('LookingForAnotherJob').sum().Count.to_dict()
groupDf['Total'] = pd.Series([total[t] for t in groupDf.LookingForAnotherJob])
groupDf['Percentage%'] = groupDf['Count']/groupDf['Total']*100
sns.catplot(x = 'LookingForAnotherJob',y = 'Percentage%',hue = 'JobTitle',kind =
```

Out[29]:

```
<seaborn.axisgrid.FacetGrid at 0x1a1f39ff90>
```



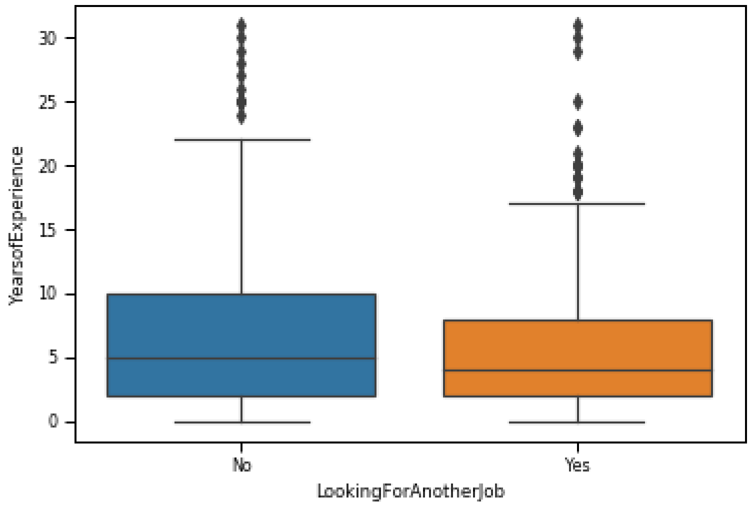
10. What are the two roles that have the highest and lowest percentage of employees looking for other jobs?

Answer from the figure, we can see that DBA (database administrator) have highest percentage of employees looking for other jobs, Data scientist have lowest employees looking for other jobs

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 [52]:

```
# Your code
datajob = df[(df.DataScienceRelated == True)]
yearOfExperience = sns.boxplot(x = 'LookingForAnotherJob', y = 'YearsofExperienc
```



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

Answer From the figure, we can find that people with less year of experience are more likely looking for another job, we assume that looking for another job means less happiness.Their average decision is about 5 years. If they work more than approximately 16 years, they are more stable, thus, more happiness.

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 [53]:

```
# Your code
df["SalaryAUD"] = df['SalaryUSD']*1.47
salary_job = df.groupby('JobTitle').SalaryAUD.describe()
salary_job = salary_job.reset_index(level=0)
salary_job = salary_job.rename(columns={"50%" : "median"})
salary_job.drop(columns=['count', 'mean', 'std', 'min', '25%','75%'])#delete ir
```

Out[53]:

	JobTitle	median	max
0	Architect	176400.0	514500.00
1	DBA	132300.0	1411200.00
2	Data Analyst	113190.0	624750.00
3	Data Engineer	139650.0	955500.00
4	Data Scientist	163170.0	235200.00
5	Developer: App code (C#, JS, etc)	117600.0	285180.00
6	Developer: T-SQL	124950.0	1036350.00
7	Manager	161700.0	924419.79

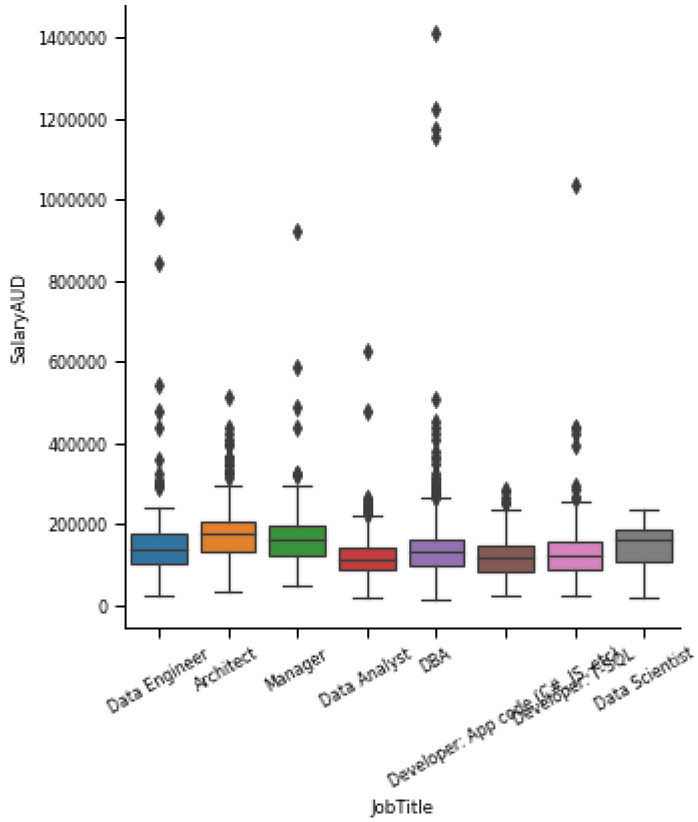
2. Do those figures confirm that data scientists are well paid?

In [156]:

```python
Job_Salary = sns.catplot(x = 'JobTitle',y = 'SalaryAUD',kind = 'box',data = df)
Job_Salary.set_xticklabels(rotation=30)#rotate the x label
```

Out[156]:

```
<seaborn.axisgrid.FacetGrid at 0x1a21c27610>
```



In [55]:

```
df.groupby('JobTitle').SalaryAUD.describe()
```

Out[55]:

JobTitle	count	mean	std	min	25%	50%	75%	
Architect	466.0	174930.492103	66746.531470	32340.00	132300.000	176400.0	205800.0	514
DBA	2997.0	134254.713184	64771.086321	16190.58	97020.000	132300.0	165154.5	1411
Data Analyst	703.0	118186.425030	48866.776508	22050.00	88200.000	113190.0	142590.0	624
Data Engineer	270.0	148883.157111	93826.877558	23520.00	102900.000	139650.0	176400.0	955
Data Scientist	48.0	148261.321250	54573.852115	17640.00	110250.000	163170.0	189630.0	235
Developer: App code (C#, JS, etc)	426.0	118304.813377	48701.762900	25902.87	83128.500	117600.0	147000.0	285
Developer: T-SQL	562.0	127666.815393	66915.124463	22785.00	90657.105	124950.0	158760.0	1036
Manager	342.0	166255.957675	75433.973622	48510.00	124398.750	161700.0	198450.0	924

Answer Yes. From the figure,we can find that its distribution is stable and likely only follow by Architect and Manager.

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.

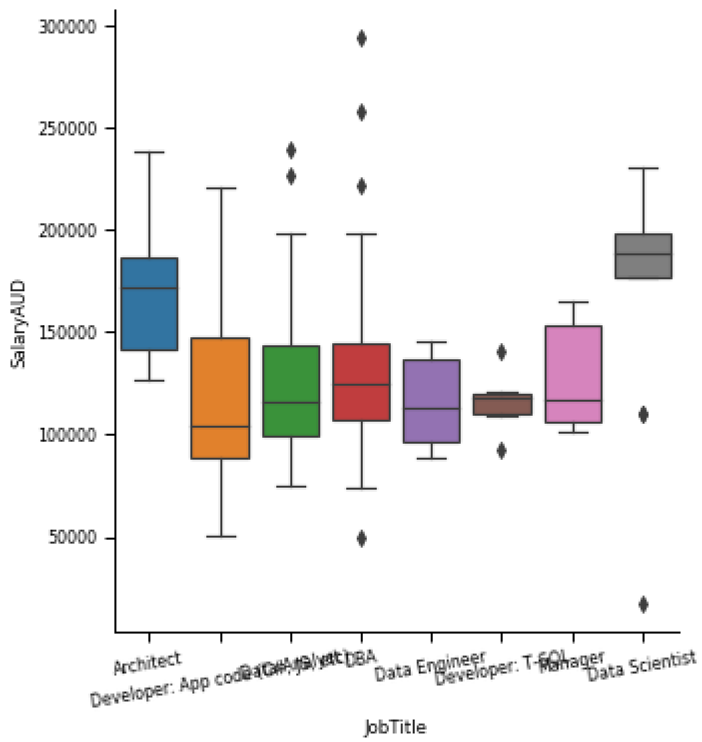
3. Plot boxplot chart of the Australian respondents salary distribution grouped by the different job titles.

In [56]:

```
# Your code
Australia = df[df.Country == 'Australia']
Australia = sns.catplot(x = 'JobTitle',y = 'SalaryAUD',kind = 'box',data = Austr
Australia.set_xticklabels(rotation=10)
```

Out[56]:

<seaborn.axisgrid.FacetGrid at 0x1a1d0bc110>



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

Answer from the figure, we can see that data scientists paid distribution is higher than other jobs.

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).

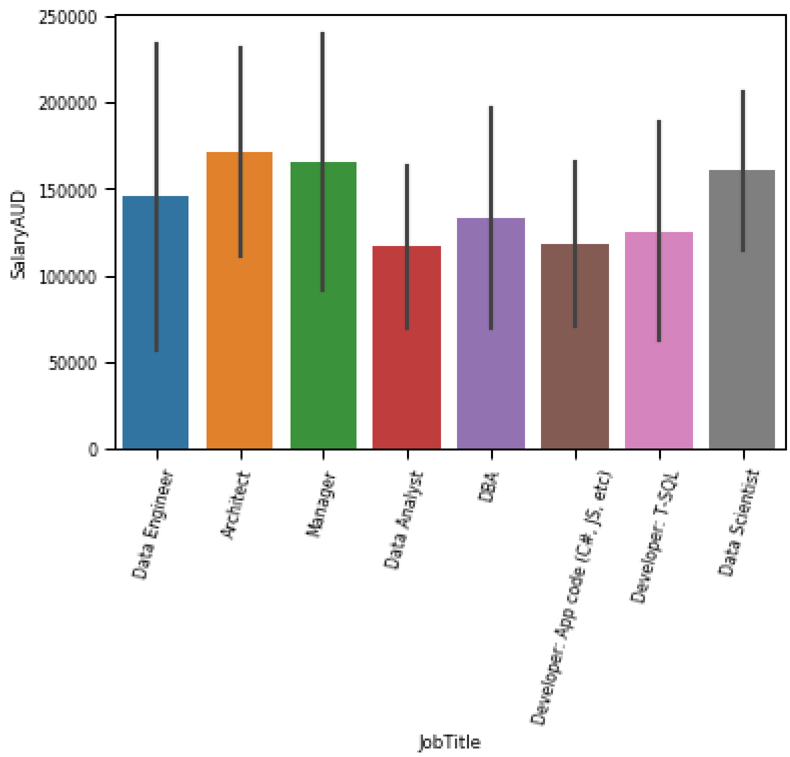
Hint: Consider all job titles and filter for full-time employees only

In [57]:

```
# Your code
allFullTime = df[(df.EmploymentStatus == 'Full time employee')]
fulltime = sns.barplot(x="JobTitle", y="SalaryAUD", data=allFullTime, ci='sd') ;
fulltime.set_xticklabels(fulltime.get_xticklabels(), rotation=75)
```

Out[57]:

```
[Text(0, 0, 'Data Engineer'),
 Text(0, 0, 'Architect'),
 Text(0, 0, 'Manager'),
 Text(0, 0, 'Data Analyst'),
 Text(0, 0, 'DBA'),
 Text(0, 0, 'Developer: App code (C#, JS, etc)'),
 Text(0, 0, 'Developer: T-SQL'),
 Text(0, 0, 'Data Scientist')]
```



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

```
<span style="color: green">**Answer**</span>
<answer> from the figure, we can find from all country that the best paid jobs are still data scientist and Architect, and the worst paid jobs are data analyst and developer. The reason maybe is that architect need more experience and better skill, data scientist is a emerging job which need math skill also solve the business problem properly. Data analyst do not need more programming skill but soft skill like communication and business background. However, market do not have that higher need for developers. Data engineer have a higher standard deviation means people in this area will have a dispersion salary.</answer>
```

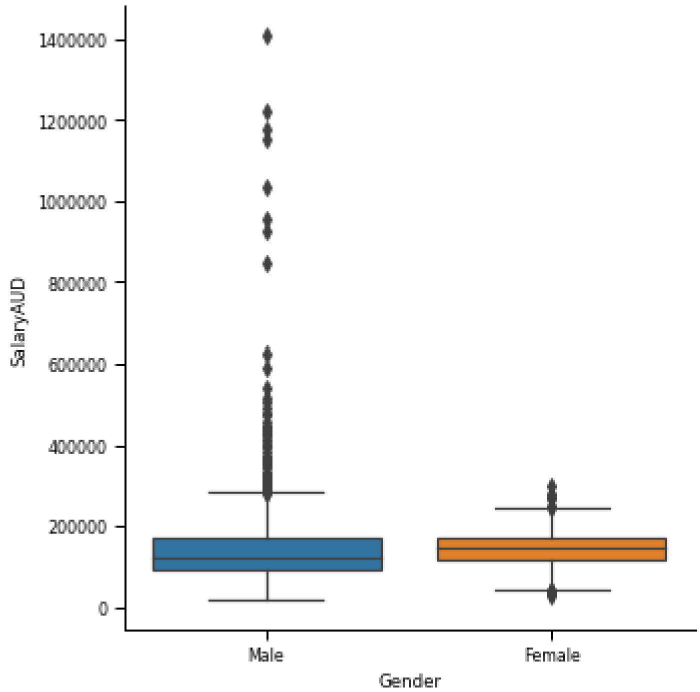
4.3 Salary and Gender

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 [58]:

```
# Your code
g = sns.catplot(x="Gender", y="SalaryAUD", kind="box",data=df);
```



8. What do you notice about the distributions?

```
<span style="color: green">**Answer**</span>
<answer>male median salary is lower than female, however, their max salary is greater than female, also, they got a lot of outlier above the max, which means that some of male got extremely high salary.</answer>
```

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 [59]:

```
# Your code
AUS_US_IND = df[(df.Country=='Australia')|(df.Country=='United States')|(df.Cour
AUS_US_IND = AUS_US_IND.groupby(['Country','Gender']).median()
AUS_US_IND.drop(columns=['Age', 'YearsofExperience', 'SalaryUSD','DataScienceRel
```

Out[59]:

		SalaryAUD
Country	**Gender**	
Australia	Female	139650.0
	Male	122010.0
India	Female	48142.5
	Male	34251.0
United States	Female	147602.7
	Male	154350.0

4.4 Salary and formal education

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

Let's see.

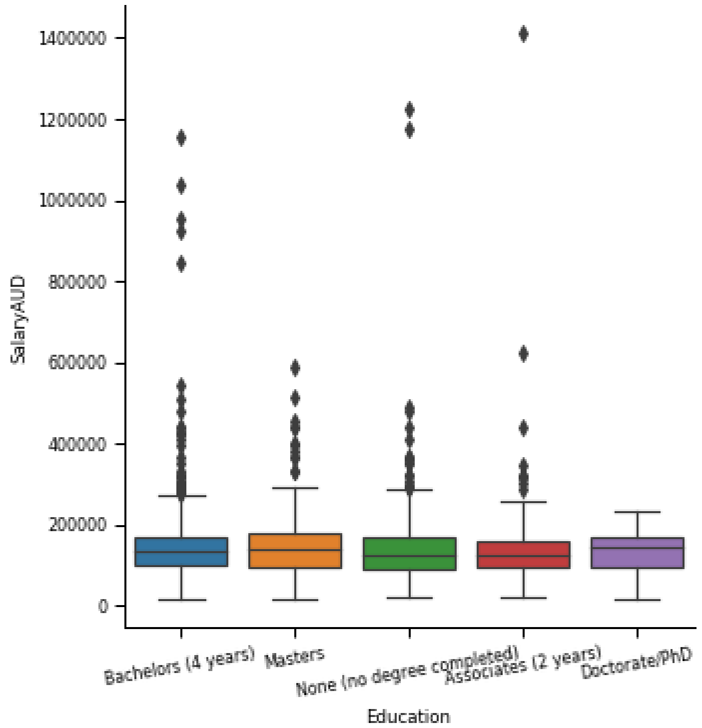
10. Plot the salary distribution of all respondants and group by formal education type on a boxplot.

In [60]:

```
# Your code
salaryEducation = sns.catplot(kind = 'box',x = 'Education',y ='SalaryAUD', data
salaryEducation.set_xticklabels(rotation=10)
```

Out[60]:

<seaborn.axisgrid.FacetGrid at 0x1a1d5d3cd0>



In [61]:

```
education_salary = df.groupby(['Education']).describe()
education_salary['SalaryAUD']
```

Out[61]:

Education	count	mean	std	min	25%	50%	75%	
Associates (2 years)	633.0	130329.108415	74256.949245	22785.00	92610.0	124950.0	158760.0	141
Bachelors (4 years)	3094.0	138071.843377	63429.112783	16190.58	99960.0	134431.5	169050.0	115
Doctorate/PhD	55.0	132684.578727	53576.313729	17640.00	94447.5	142590.0	169785.0	230
Masters	1043.0	137737.785101	62546.490485	17934.00	95550.0	136710.0	176400.0	588
None (no degree completed)	989.0	134109.610637	75019.772299	22050.00	90846.0	124950.0	169050.0	122

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

Answer Actually, from the figure, we find that there are no huge difference between degrees. Bachelors got avarage 138071.843377$, however they have many outliers. Master's 75% is slightly greater than bachelors'. For doctorate/PhD, their 75% is slightly less than masters', which is 169785. However, they have least standard deviation, 53576.313729, which means their incomes are more stable.

4.5 Salary and Employment Sector

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

Let's see.

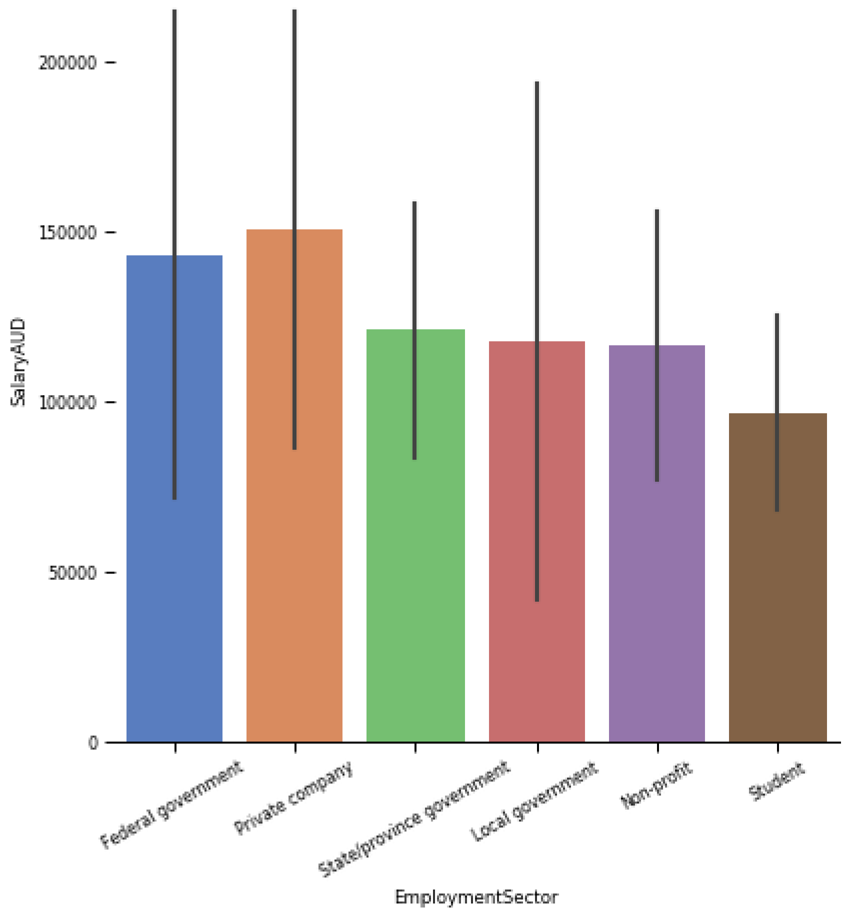
12. Plot a bar chart (with error bars) of the salaries of respondents for each of the employment sectors.

In [62]:

```
# Your code
salaryEmployment = sns.catplot(x="EmploymentSector", y="SalaryAUD", data=df,
                    height=6, kind="bar",ci='sd', palette="muted")
salaryEmployment.despine(left=True)
salaryEmployment.set_xticklabels(rotation=30)
```

Out[62]:

```
<seaborn.axisgrid.FacetGrid at 0x1a1db0ab90>
```



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 It seems like private company have the highest salary. I think it would differ based on the country. We can use hue parameter with character "Country", to demonstrate the figure. However, with so many countries in our dataset, we cannot see the trend clearly.

5. Predicting salary

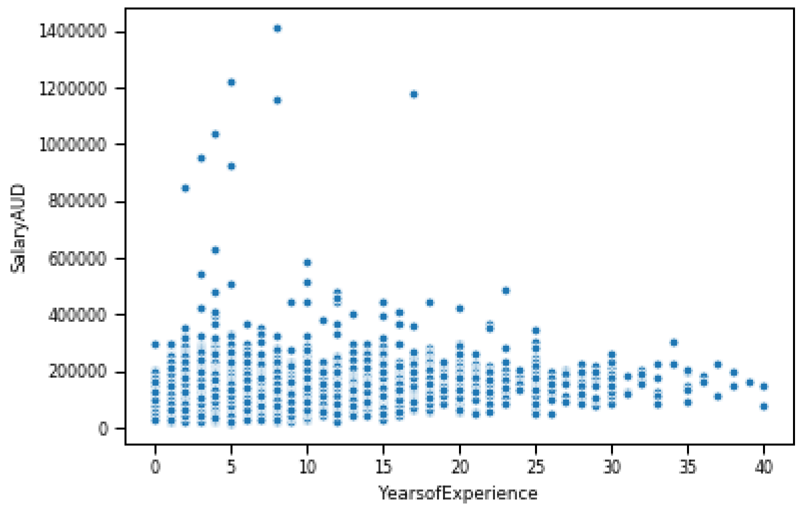
We have looked at many variables and seen that there are a lot of factors that could affect your salary.

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 [63]:

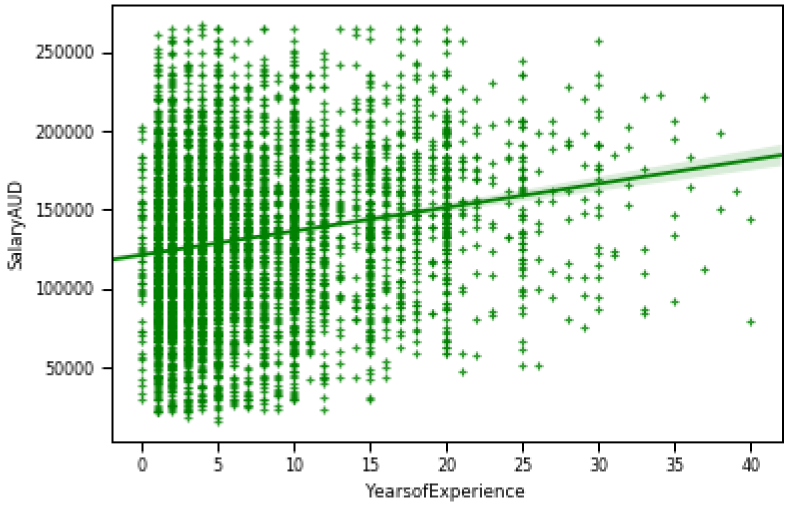
```
# Your code
salary_experience = sns.scatterplot(x = 'YearsofExperience', y = 'SalaryAUD',dat
```



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 [64]:

```
#Your code
standDeviation = df['SalaryAUD'].std()
mean = df['SalaryAUD'].mean()
removedData = df[(df['SalaryAUD'] > mean - 2*standDeviation)&(df['SalaryAUD'] <
sns.regplot(x="YearsofExperience", y="SalaryAUD", data=removedData,marker="+",co
```



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

Answer The shadow green part try to represent 95% confidence interval. However, it seems most of the points are not inside these area.For this prediction, r-square is a kind of benchmark indicating for regression accuracy.

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 [65]:

```
import numpy as np
import pandas as pd
from os import path
from PIL import Image
!pip install wordcloud #install wordcloud
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
import matplotlib.pyplot as plt
```

```
Requirement already satisfied: wordcloud in /opt/anaconda3/lib/python
3.7/site-packages (1.6.0)
Requirement already satisfied: pillow in /opt/anaconda3/lib/python3.7/
site-packages (from wordcloud) (6.2.0)
Requirement already satisfied: matplotlib in /opt/anaconda3/lib/python
3.7/site-packages (from wordcloud) (3.1.1)
Requirement already satisfied: numpy>=1.6.1 in /opt/anaconda3/lib/pyth
on3.7/site-packages (from wordcloud) (1.17.2)
Requirement already satisfied: cycler>=0.10 in /opt/anaconda3/lib/pyth
on3.7/site-packages (from matplotlib->wordcloud) (0.10.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /opt/anaconda3/li
b/python3.7/site-packages (from matplotlib->wordcloud) (1.1.0)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.
1 in /opt/anaconda3/lib/python3.7/site-packages (from matplotlib->word
cloud) (2.4.2)
Requirement already satisfied: python-dateutil>=2.1 in /opt/anaconda3/
lib/python3.7/site-packages (from matplotlib->wordcloud) (2.8.0)
Requirement already satisfied: six in /opt/anaconda3/lib/python3.7/sit
e-packages (from cycler>=0.10->matplotlib->wordcloud) (1.12.0)
Requirement already satisfied: setuptools in /opt/anaconda3/lib/python
3.7/site-packages (from kiwisolver>=1.0.1->matplotlib->wordcloud) (41.
4.0)
```

In [144]:

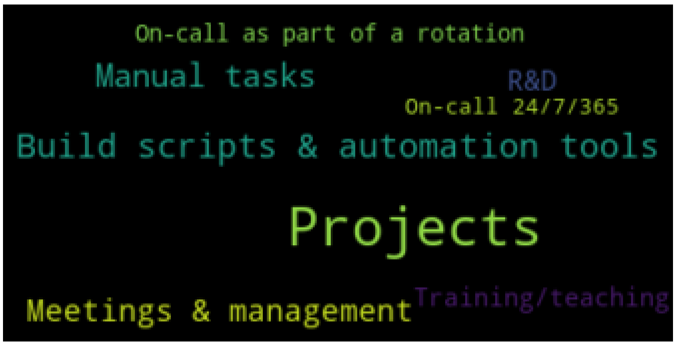
```
# Your code
splitKind = df['KindsOfTasksPerformed'].str.split(", ",expand = True)
text = pd.concat([splitKind, df['DataScienceRelated']],axis = 1)
text = text.melt(id_vars=['DataScienceRelated'], value_vars=[0,1,2,3,4,5,6,7])
text = text[~text.value.isna()]
text = text.groupby(['value']).count()
# text = text.rename(columns={"DataScienceRelated" : "Count"})
text = text.reset_index(level=0)
text2 = text.drop(columns=['variable'])

```

pandas.Series.str.split (2020) in pandas 0.25.3 documentation retrieved from
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.str.split.html
(https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.str.split.html)

In [145]:

```
d = {}
for a, x in text2.values:
    d[a] = x
import matplotlib.pyplot as plt
from wordcloud import WordCloud
wordcloud = WordCloud()
wordcloud.generate_from_frequencies(frequencies=d)
plt.figure()
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```



Ricardo.M (2020) WordCloud from data frame with frequency python retrieved from [https://stackoverflow.com/questions/38465478/wordcloud-from-data-frame-with-frequency-python (https://stackoverflow.com/questions/38465478/wordcloud-from-data-frame-with-frequency-python)](https://stackoverflow.com/questions/38465478/wordcloud-from-data-frame-with-frequency-python)

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 [152]:

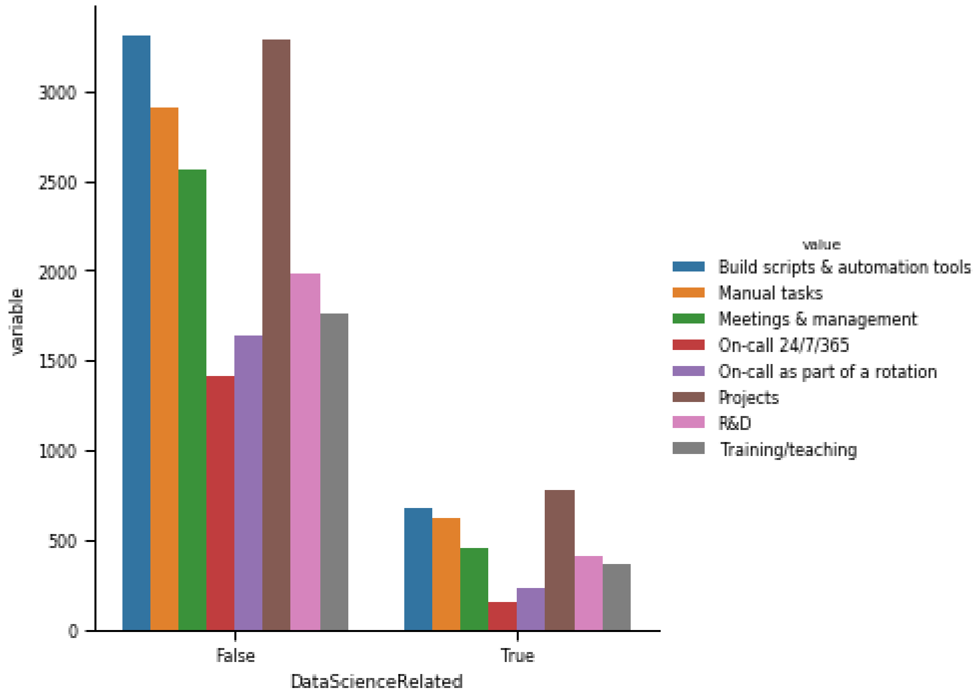
```
splitKind = df['KindsOfTasksPerformed'].str.split(", ",expand = True)
text = pd.concat([splitKind, df['DataScienceRelated']],axis = 1)
text = text.melt(id_vars=['DataScienceRelated'], value_vars=[0,1,2,3,4,5,6,7])
text = text[~text.value.isna()]
text = text.reset_index(level=0)
```

In [153]:

```
# Your code
text2 = text.groupby(['value','DataScienceRelated']).count()
text2.columns
text2 = text2.reset_index(level=1)
text2 = text2.reset_index(level=0)
text2
sns.catplot(x = 'DataScienceRelated' ,y= 'variable' ,hue = 'value',kind = 'bar',
```

Out[153]:

```
<seaborn.axisgrid.FacetGrid at 0x1a218a9910>
```



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

Answer from the image, we can see that "build scripts & automation tools" and "Projects" are the most commonly used.

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 When I do Employment 3.1.5, the distribution of full time employee is anomaly. The reason why its anomalies is it have a lot of outliers over 40 years old, which is very different from others. The reason why they may exist maybe because people work over 40 will retired. However, from google, USA average retire years is 66 years old. Thus, another reason for this question is the bias sampling. Maybe the sampling age from Q1-Q3 are not get properly. wishes to discard them or use statistics that are robust to outliers, may sometimes been discarded to ensure the robust.To fix this problem, maybe we can use other more reasonable sample method. Perhaps stratified random sampling is a good choice.

For 3.2 Job Satisfaction we observed that DBA has the highest percentage of looking for another job. However, from money_usnews job satisfaction faction, DBA got low stress level, good work-life balance and solid prospects. Thus, DBA should have a hihger job satisfaction with low percentage of looking for another job. The reason of anomaly maybe because they are actually collected by the system,without proper sample selecting method. Also, stratified random can be used.

Database Administrator (2020, January). in USNEWS money Retrieved from [https://money.usnews.com/careers/best-jobs/database-administrator (https://money.usnews.com/careers/best-jobs/database-administrator)](https://money.usnews.com/careers/best-jobs/database-administrator)

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.