PuppalaSucharitha_Assignment_2.2

September 10, 2022

0.0.1 WEEK 2

0.0.2 Assignment 2.2 - Graph Analysis with matplolib

0.0.3 Puppala Sucharitha

0.0.4 Date : 09/09/2022

Importing the necessary Libraries.

[1]:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib as mpl
import matplotlib.mlab as mlab
import matplotlib.pyplot as plt
%matplotlib inline
```

0.0.5 1 . Using a data set of your choice, write an introduction explaining the data set.

Cybersecurity is the practice of protecting systems, networks, and programs from digital attacks. These cyberattacks are usually aimed at accessing, changing, or destroying sensitive information; extorting money from users; or interrupting normal business processes. Cybersecurity is a term used to describe the process of preserving sensitive information on the internet and devices from attack, deletion, or illegal access. The cyber security goal is to provide a risk-free and secure environment in which data, networks, and devices can be protected from cyberattacks.It is a complex field, and many roles can be found within banks, retailers, e-tailers, healthcare, and government organizations. On the job, you can expect to safeguard an organization's files and network, install firewalls, create security plans, guard customer data, and monitor activity.

Here in the data set we have many different Cyber Security job roles details from the entry level and to the top level. All the data collected is from the years 2020 to 2022. The data set has 1247 rows and 11 variables in the columns. The following are the columns in the data set : work_year,experience_level,employment_type,job_title,salary,salary_currency,salary_in_usd, employee_residence,remote_ratio,company_location,company_size.

[2]:

```
# Load the file Cyber Security Salaries file.

cyberdf = pd.read_csv("salaries_cyber.csv")
```

[3]:

```
# Getting the first five rows of the data set.
cyberdf.head(5)
```

[3]:

```
   work_year experience_level employment_type               job_title  salary  \
0       2022               EN              FT    Cyber Program Manager   63000
1       2022               MI              FT         Security Analyst   95000
2       2022               MI              FT         Security Analyst   70000
3       2022               MI              FT      IT Security Analyst  250000
4       2022               EN              CT   Cyber Security Analyst  120000

  salary_currency  salary_in_usd employee_residence  remote_ratio  \
0             USD          63000                 US            50
1             USD          95000                 US             0
2             USD          70000                 US             0
3             BRL          48853                 BR            50
4             USD         120000                 BW           100

  company_location company_size
0               US            S
1               US            M
2               US            M
3               BR            L
4               BW            S
```

[4]:

```
# To get the shape of the data set i.e. how many rows and how many columns.
cyberdf.shape
```

[4]:

```
(1247, 11)
```

[5]:

```
# To get the size of the data set.
cyberdf.size
```

[5]:

```
13717
```

[6]:

```
# Information of the data set variables.
cyberdf.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1247 entries, 0 to 1246
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   work_year           1247 non-null   int64
 1   experience_level    1247 non-null   object
 2   employment_type     1247 non-null   object
 3   job_title           1247 non-null   object
 4   salary              1247 non-null   int64
 5   salary_currency     1247 non-null   object
 6   salary_in_usd       1247 non-null   int64
 7   employee_residence  1247 non-null   object
 8   remote_ratio        1247 non-null   int64
 9   company_location    1247 non-null   object
 10  company_size        1247 non-null   object
dtypes: int64(4), object(7)
memory usage: 107.3+ KB
```

Data Cleaniing.

[7]:
```
# Checking for Null Values from the data set.
cyberdf.isna().sum()
```

[7]:
```
work_year             0
experience_level      0
employment_type       0
job_title             0
salary                0
salary_currency       0
salary_in_usd         0
employee_residence    0
remote_ratio          0
company_location      0
company_size          0
dtype: int64
```

[8]:
```
# Checking for duplicates.
cyberdf.duplicated()
```

[8]:
```
0       False
1       False
2       False
3       False
4       False
        …
1242    False
1243    False
1244    False
1245    False
1246    False
Length: 1247, dtype: bool
```

[9]:
```
# getting the count of duplicates.
cyberdf.duplicated().sum()
```

[9]:
```
85
```

[10]:
```
# Drop the duplicates.
cyberdf.drop_duplicates()
```

[10]:

```
      work_year experience_level employment_type  \
0          2022               EN              FT   
1          2022               MI              FT   
2          2022               MI              FT   
3          2022               MI              FT   
4          2022               EN              CT   
…         …              …             …   
1242       2020               MI              FT   
1243       2021               SE              FT   
1244       2021               SE              FT   
1245       2021               MI              FT   
1246       2021               MI              FT   

                         job_title  salary salary_currency  salary_in_usd  \
0            Cyber Program Manager   63000             USD          63000   
1                 Security Analyst   95000             USD          95000   
2                 Security Analyst   70000             USD          70000   
3              IT Security Analyst  250000             BRL          48853   
4           Cyber Security Analyst  120000             USD         120000   
…                            …     …             …            …   
1242        Cyber Security Analyst  140000             AUD          96422   
1243  Information Security Manager   60000             GBP          82528   
1244  Penetration Testing Engineer  126000             USD         126000   
1245  Information Security Analyst   42000             GBP          57769   
1246   Threat Intelligence Analyst   66310             USD          66310   

     employee_residence  remote_ratio company_location company_size  
0                    US            50               US            S  
1                    US             0               US            M  
2                    US             0               US            M  
3                    BR            50               BR            L  
4                    BW           100               BW            S  
…                  …           …              …          …  
1242                 AU            50               AU            M  
1243                 GB            50               GB            L  
1244                 US           100               US            L  
1245                 GB           100               GB            L  
1246                 US             0               US            L  

[1162 rows x 11 columns]
```

[11]:

```
# Describing the data set.
cyberdf.describe()
```

[11]:

```
         work_year        salary  salary_in_usd  remote_ratio
count  1247.000000  1.247000e+03    1247.000000   1247.000000
mean   2021.316760  5.608525e+05  120278.218925     71.491580
```

```
std       0.715501  1.415944e+07   70291.394942    39.346851
min    2020.000000  1.740000e+03    2000.000000     0.000000
25%    2021.000000  7.975450e+04   74594.500000    50.000000
50%    2021.000000  1.200000e+05  110000.000000   100.000000
75%    2022.000000  1.600800e+05  150000.000000   100.000000
max    2022.000000  5.000000e+08  910991.000000   100.000000
```

0.0.6 2. Identify a question or question(s) that you would like to explore in your data set.

From the analysis of the cyber security slaries data set I would like to explore the following questions. 1. From the data set which level of experience has more salaries. 2. What are the top ten Job roles in the Cyber Security field? 3. Which countries has the highest Company Locations and Employee Residency? 4. Impact of Company size on salary in USD. 5. Which year is the most work year?

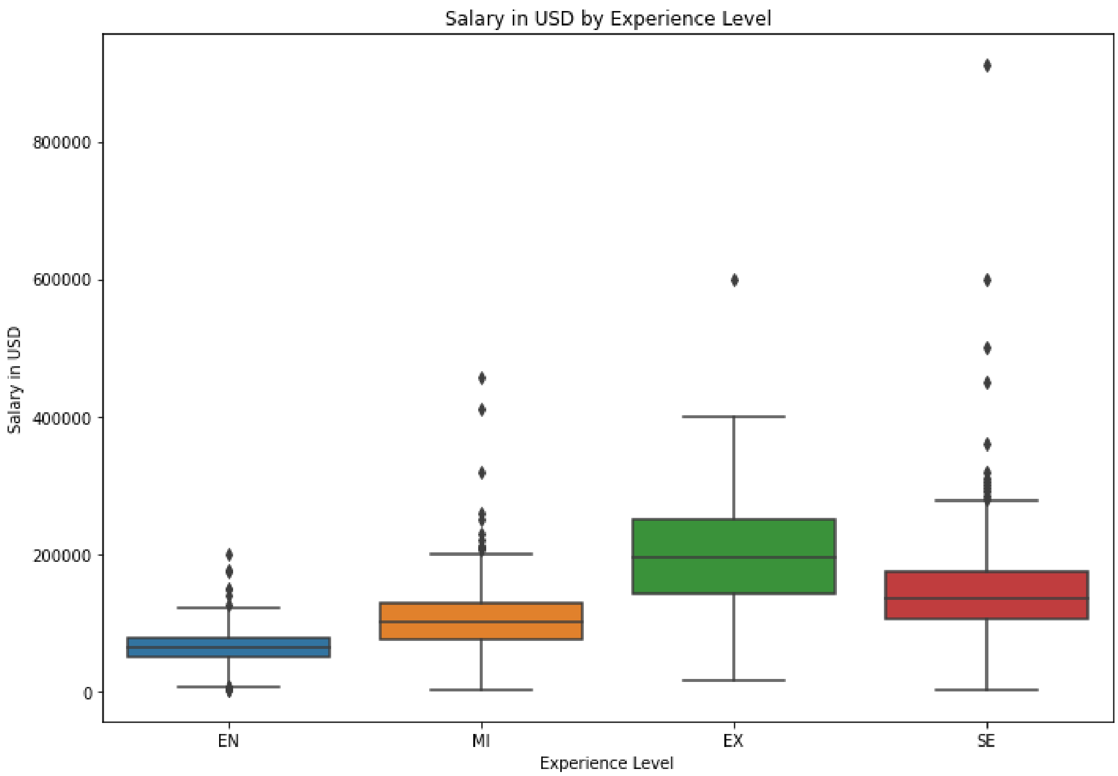
0.0.7 3. Create at least three graphs that help answer these questions. Make sure your graphs are clearly readable and are labeled appropriately and professionally.

[12]:

```python
# Box plot for Experience level and Salary.
fig, ax = plt.subplots(1, 1, figsize=(10,7), tight_layout = True)
sns.boxplot(x='experience_level', y='salary_in_usd', data=cyberdf)
plt.xlabel('Experience Level')
plt.ylabel('Salary in USD')
plt.title('Salary in USD by Experience Level')
```

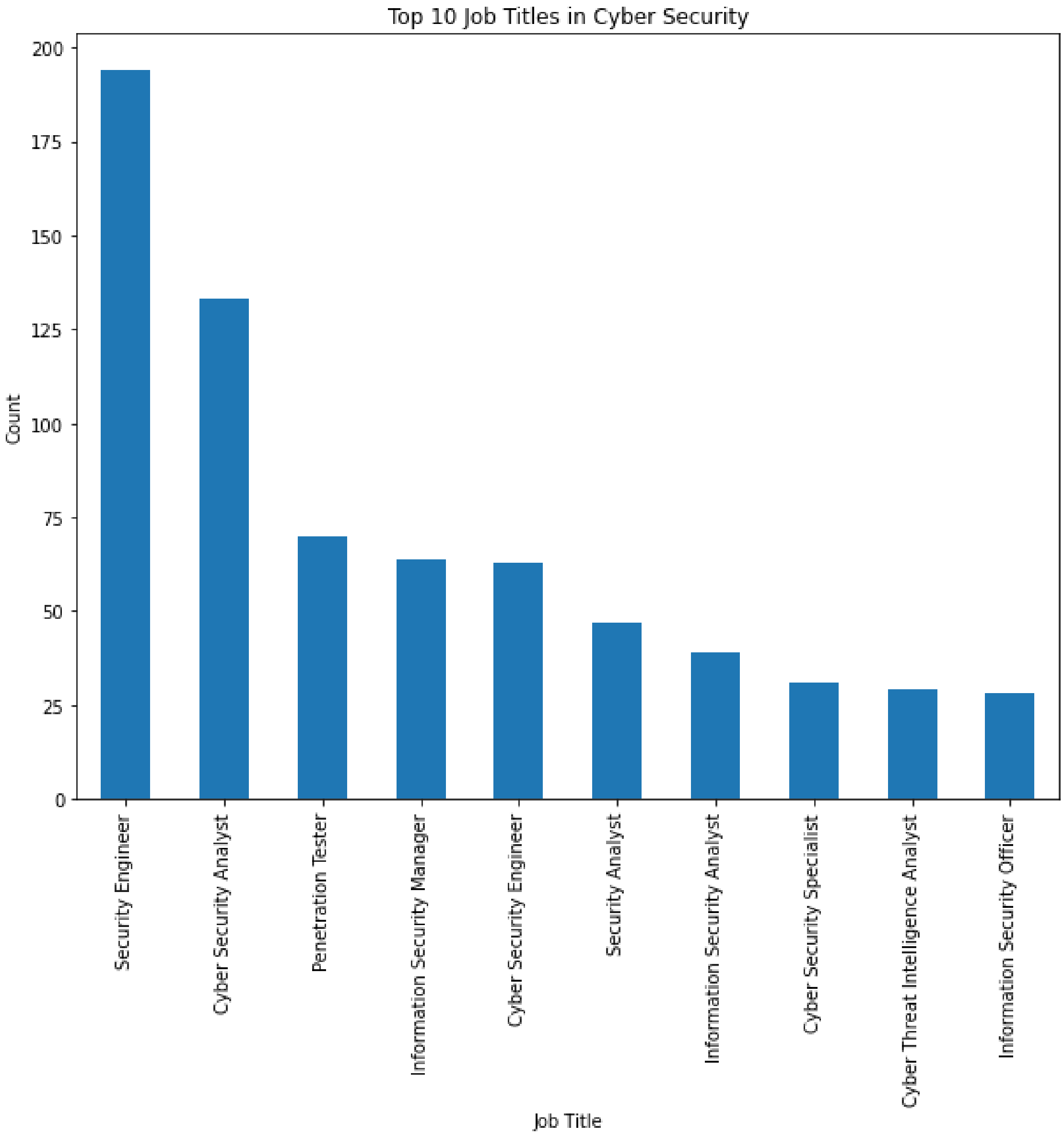
[12]:

```
Text(0.5, 1.0, 'Salary in USD by Experience Level')
```



```
[13]: # Bar plot to get the top 10 job titles in Cyber Security.
      top_10 = cyberdf['job_title'].value_counts()[:10]
      top_10.plot(kind='bar',figsize=(10,8))
      plt.title('Top 10 Job Titles in Cyber Security')
      plt.xlabel('Job Title')
      plt.ylabel('Count')
```

```
[13]: Text(0, 0.5, 'Count')
```

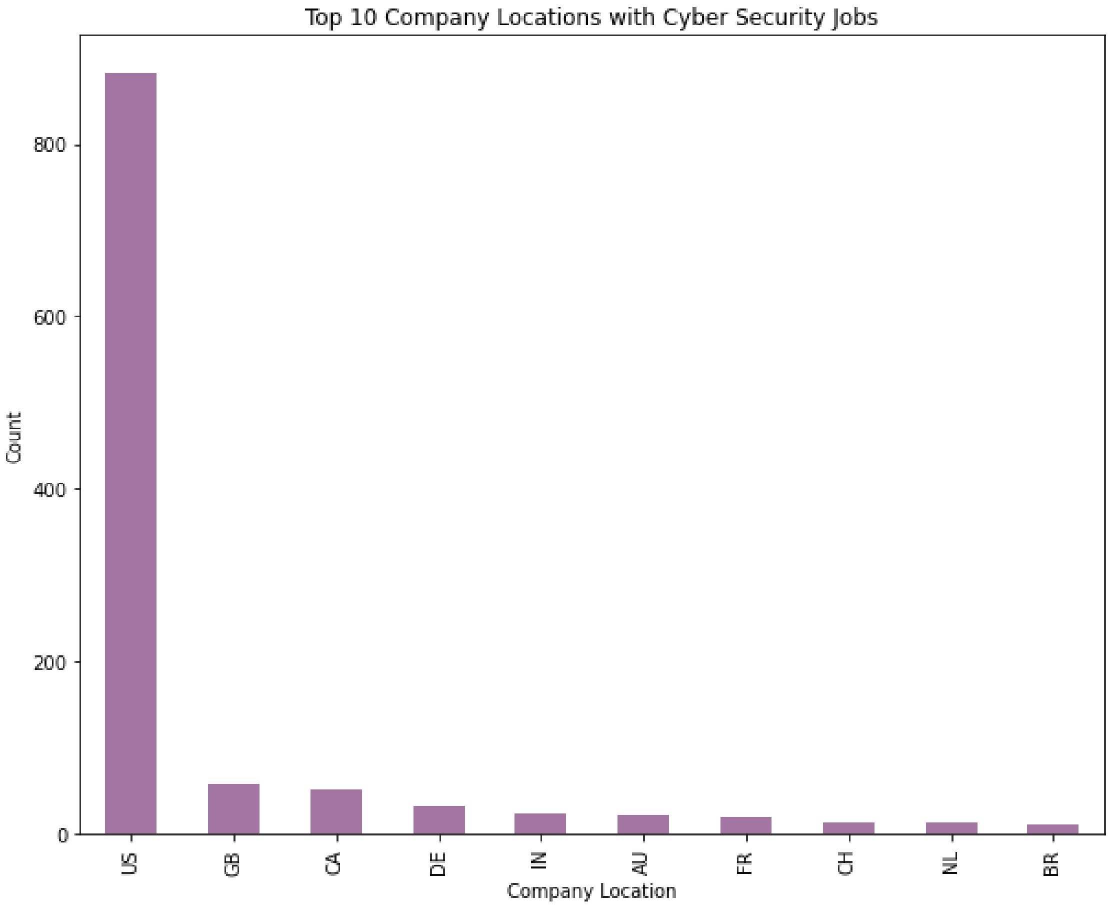


[14]:

```
# Bar plot to get the top 10 Company locations with Cyber Security jobs
top_10 = cyberdf['company_location'].value_counts()[:10]
top_10.plot(kind='bar',figsize=(10,8),color = (0.4,0.1,0.4,0.6))
plt.title('Top 10 Company Locations with Cyber Security Jobs')
plt.xlabel('Company Location')
plt.ylabel('Count')
```

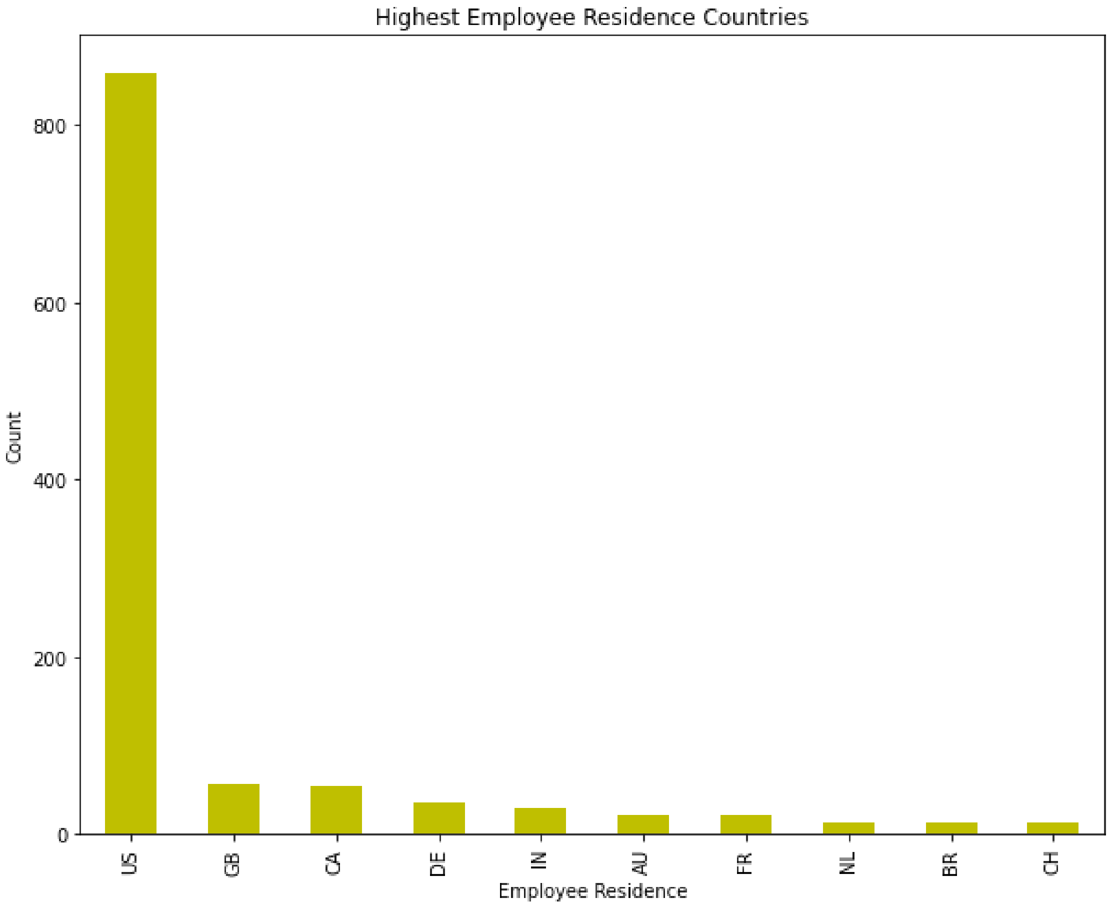
[14]:

```
Text(0, 0.5, 'Count')
```



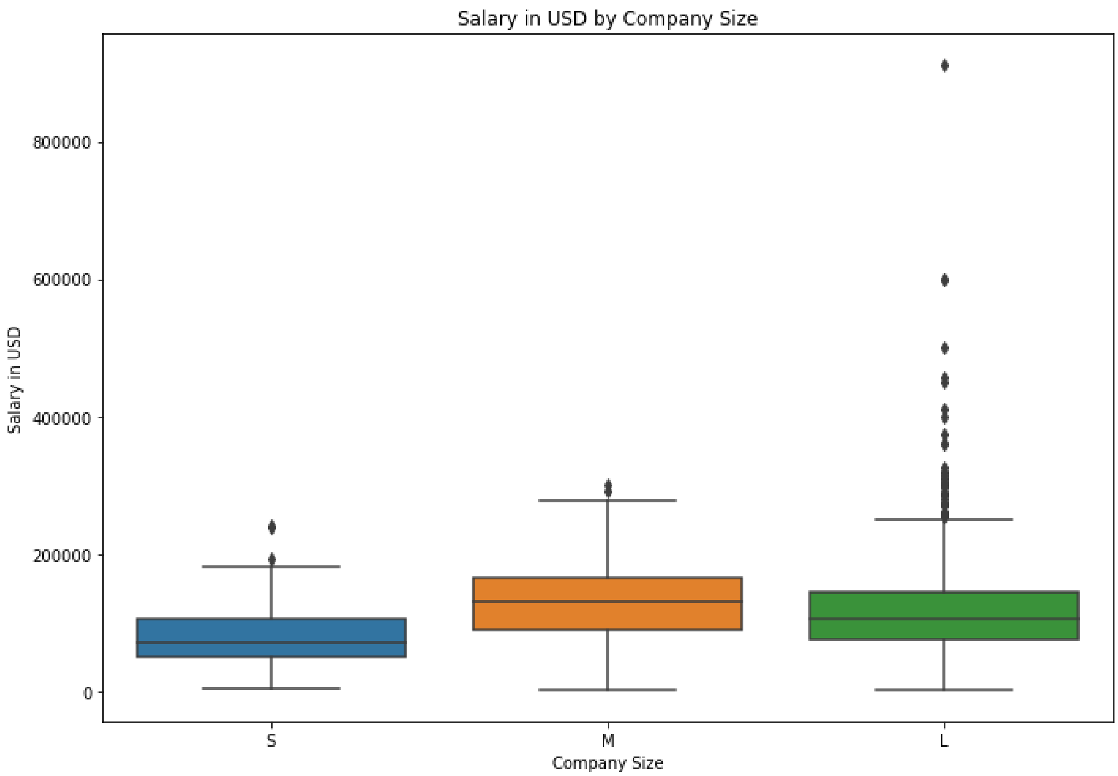
```
[15]: # Bar plot to get which country is having highest employee residence
      top_10 = cyberdf['employee_residence'].value_counts()[:10]
      top_10.plot(kind='bar',figsize=(10,8), color = 'y')
      plt.title('Highest Employee Residence Countries')
      plt.xlabel('Employee Residence')
      plt.ylabel('Count')
```

```
[15]: Text(0, 0.5, 'Count')
```



[16]:
```
# Boxplot for Company size and salary in USD.
fig, ax = plt.subplots(1, 1, figsize=(10,7), tight_layout = True)
sns.boxplot(x='company_size', y='salary_in_usd', data=cyberdf)
plt.xlabel('Company Size')
plt.ylabel('Salary in USD')
plt.title('Salary in USD by Company Size')
```

[16]:
```
Text(0.5, 1.0, 'Salary in USD by Company Size')
```

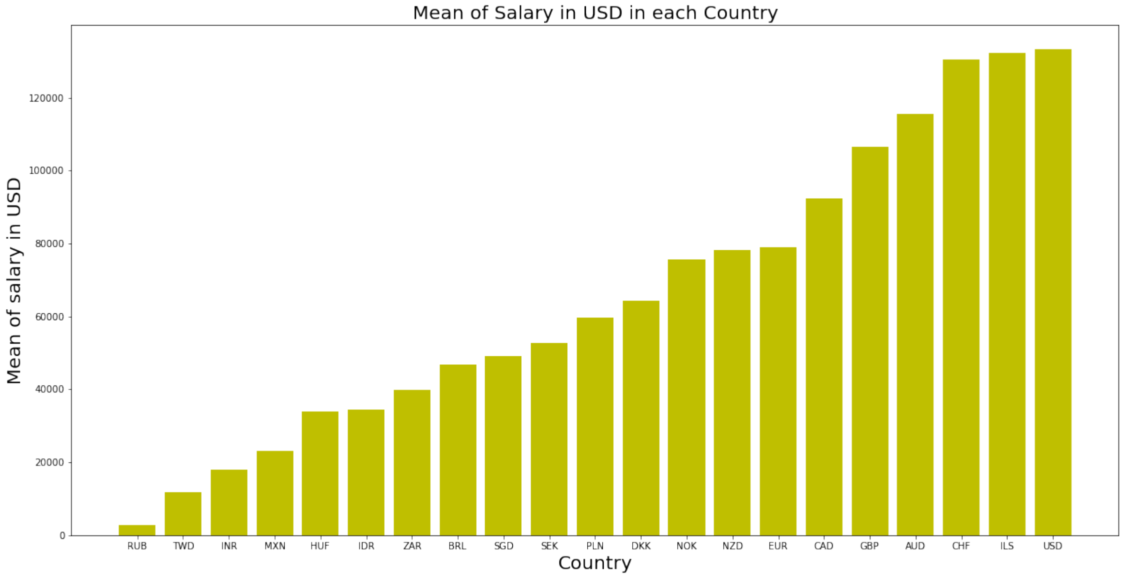


[17]:

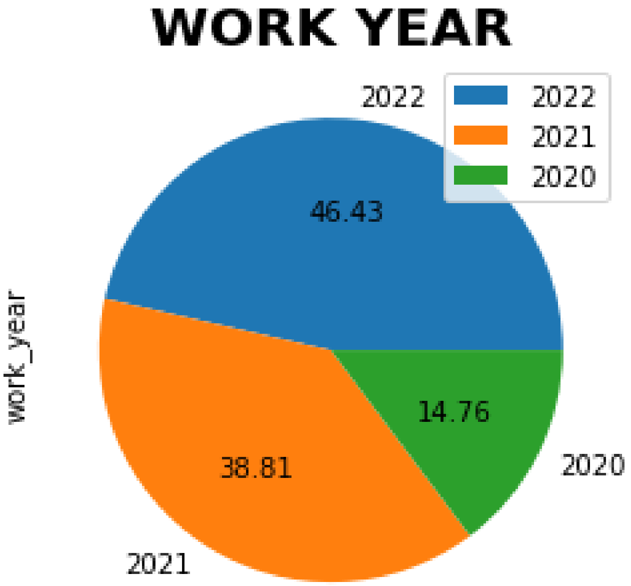
```
# To know which country people earn more salary in usd in the cyber security␣
 ↪jobs.
# As it is hard to know about the salaries in usd for all the countries,
# Consider the mean of the salaries in usd of all the countries and plot the␣
 ↪result.
saldat = cyberdf.groupby(cyberdf['salary_currency'])['salary_in_usd'].mean()
salusd = pd.DataFrame({'country':saldat.index, 'mean_usd':saldat.values})
salusd=salusd.sort_values(by=['mean_usd'])
```

[18]:

```
# Plotting bar plot for the mean of salaries in usd by countries.
a=salusd['country']
b=salusd['mean_usd']
fig = plt.figure(figsize =(20,10))
plt.bar(a,b,color='y')
plt.xlabel("Country",fontsize = 20, )
plt.ylabel("Mean of salary in USD",fontsize = 20)
plt.title("Mean of Salary in USD in each Country",fontsize = 20)
plt.show()
```

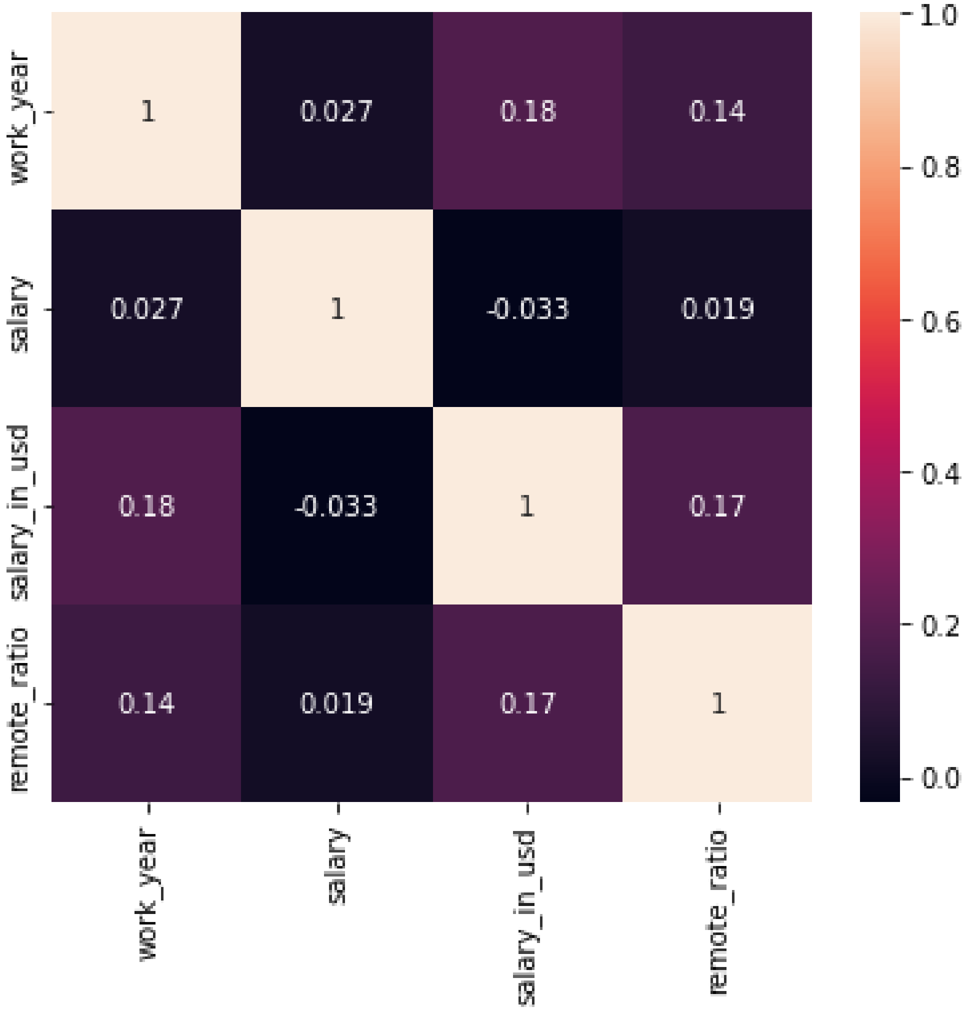


[19]:
```
# Pie plot to know which year has highest work.
labels=['2022','2021','2020']
cyberdf['work_year'].value_counts().plot(kind="pie",labels=labels,autopct="%.2f")
plt.title("WORK YEAR",fontsize = 20, fontweight='bold')
plt.legend()
plt.show()
```



[20]:

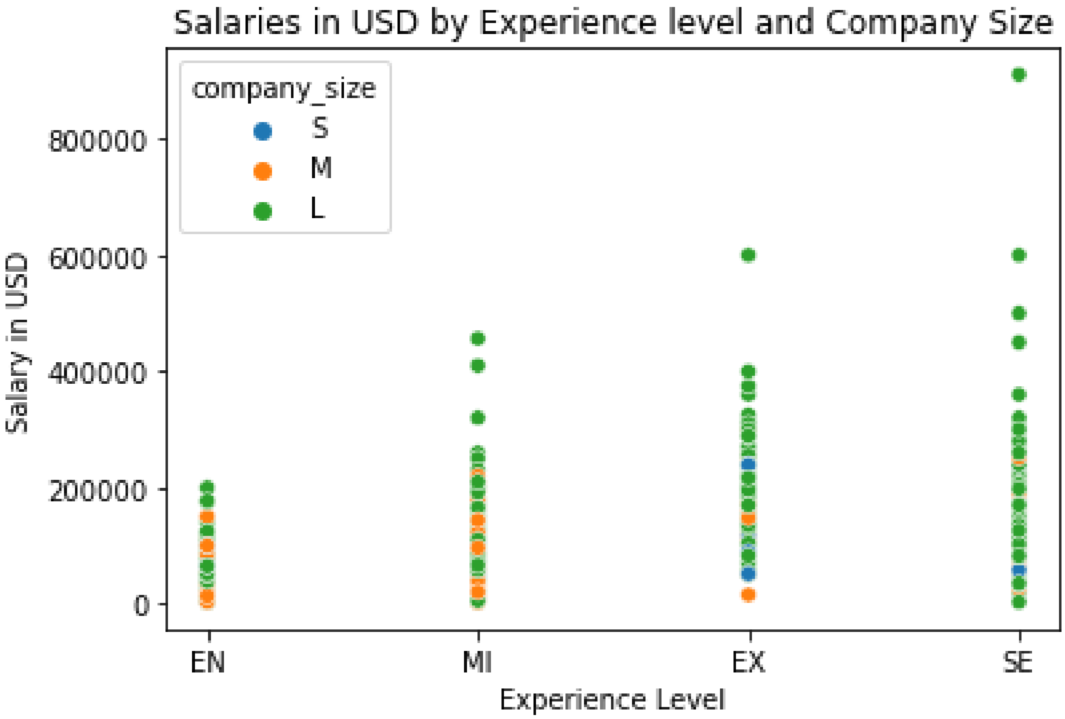
```
# Correlation heatmap for the data set.
f = plt.figure(figsize=(5.5, 5.5))
f.patch.set_facecolor('w')
sns.heatmap(cyberdf.corr(), annot=True)
plt.tight_layout()  # auto-adjust margins
```



[21]:

```
# Scatter plot to know highest salaries in USD by experience level and company␣
↪size.
sns.scatterplot(data=cyberdf, x="experience_level", y="salary_in_usd",␣
↪hue="company_size")
plt.xlabel('Experience Level')
plt.ylabel('Salary in USD')
plt.title('Salaries in USD by Experience level and Company Size')
```

```
[21]: Text(0.5, 1.0, 'Salaries in USD by Experience level and Company Size')
```



0.0.8 4. Explain what you have learned from each of your graphs.

From the above analysis the following are the different plots that are plotted from the cyber security salaries data set.

1. Box plot for Experience level and Salary.
2. Bar plot to get the top 10 job titles in Cyber Security.
3. Bar plot to get the top 10 Company locations with Cyber Security jobs.
4. Bar plot to get which country is having highest employee residence.
5. Boxplot for Company size and salary in USD.
6. Plotting bar plot for the mean of salaries in usd by countries.
7. Pie plot to know which year has highest work.
8. Correlation heatmap for the data set.
9. Scatter plot to know highest salaries in USD by experience level and company size.

There are different job titles in the field of Cyber Security and the salries paid to each role has changed depending on the experience level of the employee, company size.The experience levels are entry level(EL),mid level(MI), executive level(EX) and senior executive level(SE).The boxplot shows that EX level in experience level are highly paid,next comes the SE level.From the bar plot the top 10 job titles are extracted where we can see Security Engineer is the most popular job title among all the job titiles in the Cyber Security.The bar plot gives that US is in the top of the list in top 10 company locations for the Cyber Security jobs. The bar plot says that US is the to country that has more employee residence who are working in cyber Security next comes Isriel,Switzerland, Australia.As the data collected is from 2020 to 2022, I have considered a pie plot to get the year the job roles are more, and from the pie plot we can say that 2022 is having the higest job roles with 46.43% when compared to the remaining years.From the correlation map we cannot get to a proper solution,to arrive at a solution whether the salary paid in USD depends on experience level , company size or not, I have plotted a scatter plot, where we can find that the people working in the large sized companies are being paid more slaries in all experience levels when compared to the medium and small sized companies.

0.0.9 5. Write a conclusion that summarizes your findings.

From all the above analysis I would like to summarize that the Cyber Security job market is huge in US. The salries paid to the different job titles in the Cyber Security is dependent on the size of the company and experience level. The large sized companies pay high slaries and the executive level of experience are paid high.US is the top county in having highest number of employee residence when compared to other countries.The job market for the Cyber Security has been increased during all the past three years for which the data is collected and can say that 2022 has the huge job market for Cyber Security. Above all, with the above analysis we can say that US has the huge job market for Cyber Security when compared to other countries.