Importing Required Libraries

In [ ]:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
```

In [10]:

```
data = pd.read_csv("Latest_Data_Science_Salaries.csv")
```

In [11]:

```
print(data.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3300 entries, 0 to 3299
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   Job Title           3300 non-null   object
 1   Employment Type     3300 non-null   object
 2   Experience Level    3300 non-null   object
 3   Expertise Level     3300 non-null   object
 4   Salary              3300 non-null   int64 
 5   Salary Currency     3300 non-null   object
 6   Company Location    3300 non-null   object
 7   Salary in USD       3300 non-null   int64 
 8   Employee Residence  3300 non-null   object
 9   Company Size        3300 non-null   object
 10  Year                3300 non-null   int64 
dtypes: int64(3), object(8)
memory usage: 283.7+ KB
None
```

In [12]:
```
print(data.head())
```

```
        Job Title Employment Type Experience Level Expertise Level  Salary  \
0   Data Engineer       Full-Time           Senior          Expert  210000   
1   Data Engineer       Full-Time           Senior          Expert  165000   
2   Data Engineer       Full-Time           Senior          Expert  185900   
3   Data Engineer       Full-Time           Senior          Expert  129300   
4  Data Scientist       Full-Time           Senior          Expert  140000   

        Salary Currency Company Location  Salary in USD Employee Residence  \
0  United States Dollar    United States         210000      United States   
1  United States Dollar    United States         165000      United States   
2  United States Dollar    United States         185900      United States   
3  United States Dollar    United States         129300      United States   
4  United States Dollar    United States         140000      United States   

  Company Size  Year  
0       Medium  2023  
1       Medium  2023  
2       Medium  2023  
3       Medium  2023  
4       Medium  2023  
```

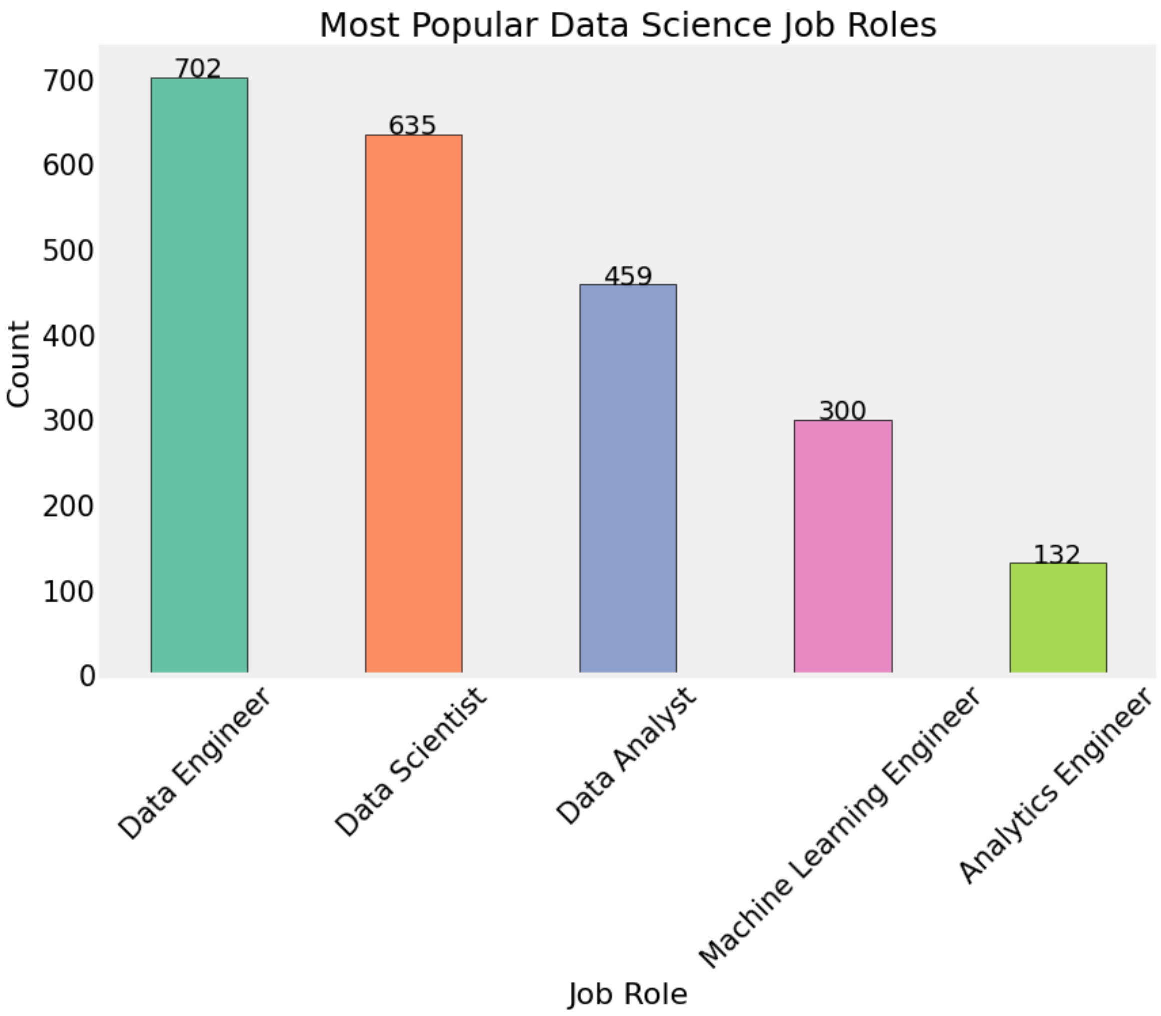
In [13]:
```
data['Year'].value_counts()
```

Out[13]:
```
Year
2023    1996
2022    1016
2021     215
2020      73
Name: count, dtype: int64
```

Most Popular Job Roles in Data Science Field

In [14]:

```
plt.figure(figsize=(10,6))
plt.style.use('fivethirtyeight')
jobtitle = data['Job Title'].value_counts()

def addlabels(x,y):
    for i in range(len(x)):
        plt.text(i, y[i], y[i], ha = 'center', fontsize=14, color='black')

plt.bar(jobtitle.index[0:5], jobtitle.values[0:5], color=sns.color_palette("Se
plt.xticks(rotation=45, fontsize=15.0, color='black')
plt.yticks(fontsize = 15.0, color='black')
addlabels(jobtitle.index[0:5], jobtitle.values[0:5])
plt.xlabel('Job Role', fontsize=16, color='black')
plt.ylabel('Count', fontsize=16, color='black')
plt.grid(False)
plt.title('Most Popular Data Science Job Roles', fontsize = 18.0, color='black

plt.show()
```

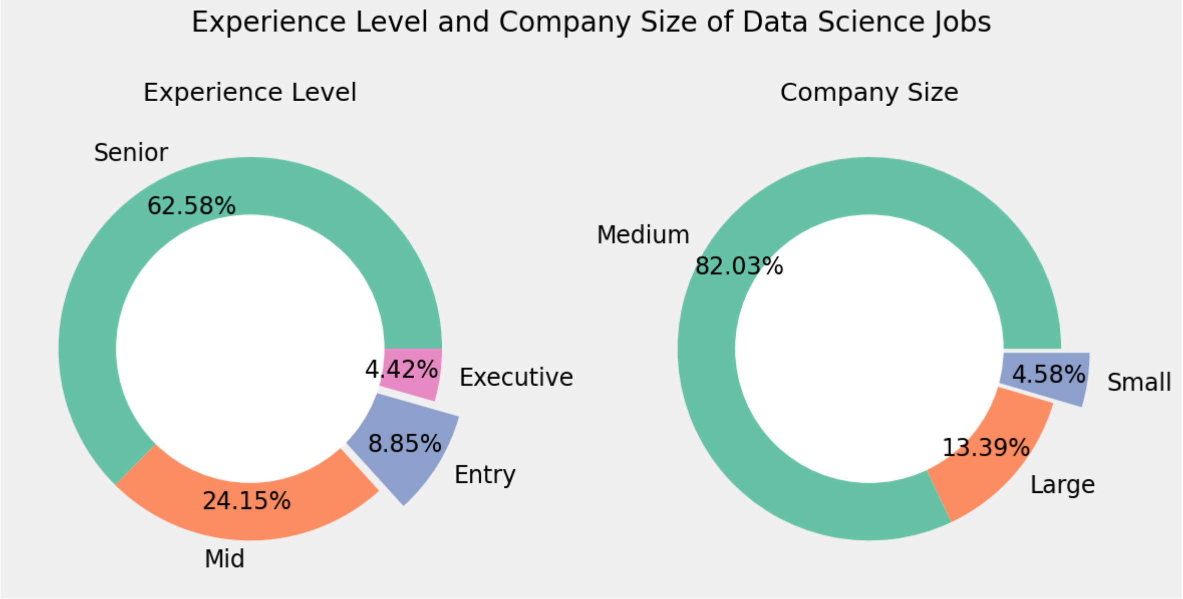


Pie Chart - Experience and Expertise Level

```
In [15]: exp_level = data['Experience Level'].value_counts()
         print(exp_level)

         expert_level = data['Expertise Level'].value_counts()
         print(expert_level)

         comp_size = data['Company Size'].value_counts()
         print(comp_size)

         comp_loc = data['Company Location'].value_counts()
         print(comp_loc)

         emp_loc = data['Employee Residence'].value_counts()
         print(emp_loc)
```

```
Experience Level
Senior       2065
Mid           797
Entry         292
Executive     146
Name: count, dtype: int64
Expertise Level
Expert          2065
Intermediate     797
Junior           292
Director         146
Name: count, dtype: int64
Company Size
Medium    2707
Large      442
Small      151
Name: count, dtype: int64
Company Location
United States             2495
United Kingdom             251
Canada                     104
Germany                     65
Spain                       47
                          ... 
Korea, Republic of           1
Armenia                      1
Andorra                      1
Bosnia and Herzegovina       1
Malta                        1
Name: count, Length: 71, dtype: int64
Employee Residence
United States                2453
United Kingdom                245
Canada                        101
Germany                        58
India                          57
                             ... 
Bosnia and Herzegovina          1
American Samoa                  1
Iran, Islamic Republic of       1
Kenya                           1
Malta                           1
Name: count, Length: 83, dtype: int64
```

In [16]:
```
plt.style.use('fivethirtyeight')
plt.figure(figsize=(12,10))
plt.suptitle('Experience Level and Company Size of Data Science Jobs',y=0.8,ha
            fontsize=20.0, color='black')

plt.subplot(1,2,1)
explode = [0,0,0.15,0]
plt.pie(exp_level, labels = exp_level.index, autopct = '%1.2f%%',
       pctdistance = 0.8, explode = explode, colors = sns.color_palette("Set2'
       textprops={'fontsize':17})

plt.title('Experience Level',fontsize=18.0, color='black')

hole = plt.Circle((0,0), 0.70, facecolor = 'white')

plt.gcf().gca().add_artist(hole)
#-------------------------
plt.subplot(1,2,2)
explode = [0,0,0.15]
plt.pie(comp_size, labels = comp_size.index, autopct = '%1.2f%%',
       pctdistance = 0.8, explode = explode, colors = sns.color_palette("Set2'
       textprops={'fontsize':17})

plt.title('Company Size',fontsize=18.0, color='black')

hole = plt.Circle((0,0), 0.70, facecolor = 'white')

plt.gcf().gca().add_artist(hole)
plt.tight_layout()
plt.show()
```



Company Location vs Employee Location

In [17]:

```
location = pd.DataFrame(data['Company Location'].value_counts())
```

In [18]:

```
location['Employee Location'] = data['Employee Residence'].value_counts()
```

In [19]:

```
location = location.rename(columns = {'count':'Number of Companies',
                                      'Employee Location':'Number of Employees'
```

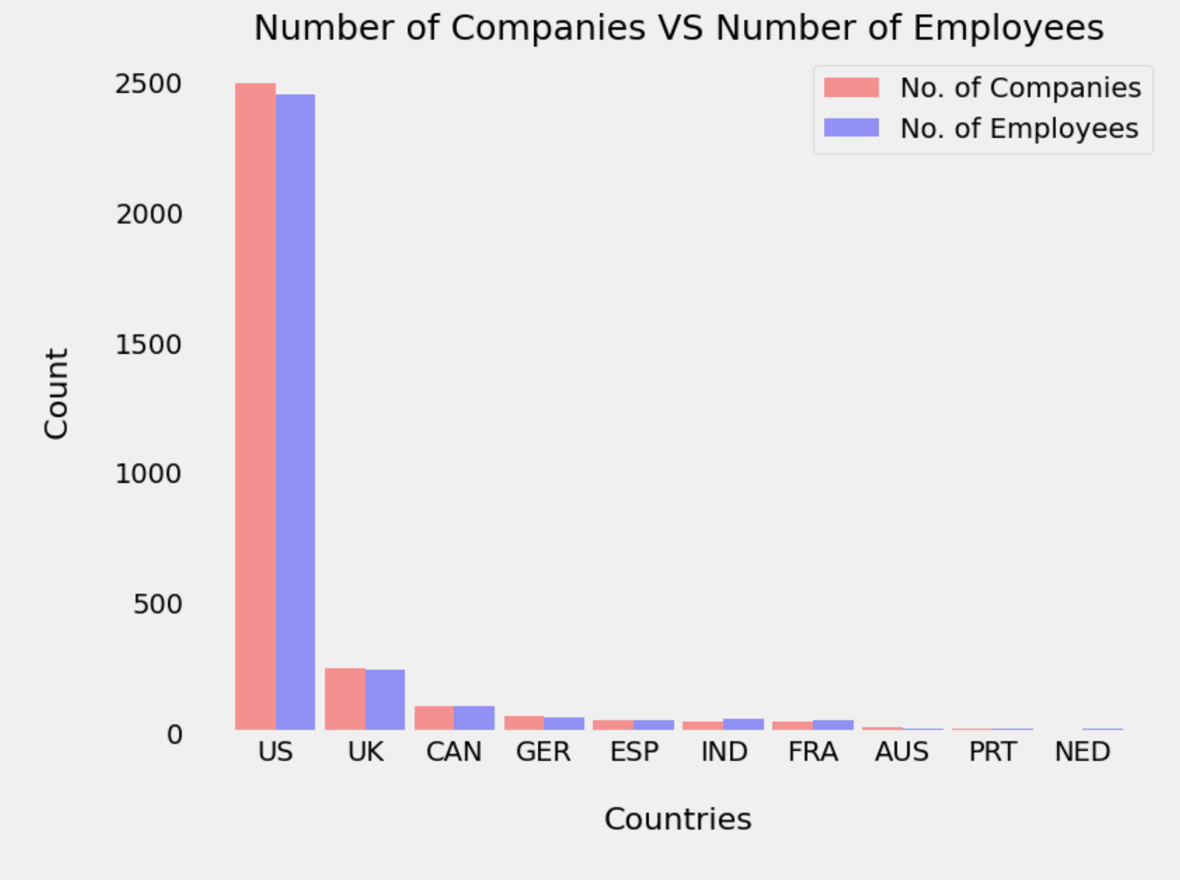
In [20]:

```
location = location.reset_index()
```

In [21]:

```
sorter = location['Company Location'][0:10].value_counts().index
```

In [22]:
```
x = np.arange(len(sorter))
width = 0.45
plt.style.use('fivethirtyeight')
fig, ax = plt.subplots(1, figsize=(8,6))
ax.bar(x - width/2, location['Number of Companies'][0:10], width, label='No. c
      color = 'red', alpha=0.4)
ax.bar(x + width/2, location['Number of Employees'][0:10], width, label='No. c
      color = 'blue', alpha=0.4)

plt.xticks([i for i in range(len(sorter))],['US', 'UK', 'CAN', 'GER', 'ESP', '
          color='black')
plt.yticks(color='black')
plt.title('Number of Companies VS Number of Employees', fontsize=18, color='bl
plt.xlabel('\nCountries\n', fontsize=16, color='black')
plt.ylabel('\n Count \n', fontsize=16, color='black')
plt.grid(False)
ax.legend(fontsize=14)

plt.show()
```



In [23]:
```
salary = data[['Job Title', 'Salary in USD', 'Year' ]]

salary = salary.groupby(['Job Title', 'Year']).mean()
```

In [24]:
```
salary = salary.reset_index()
```

In [25]:
```
salary.sort_values(by = 'Salary in USD', ascending=False)
```

Out[25]:
```
                              Job Title  Year  Salary in USD
 69                 Data Analytics Lead  2022       405000.0
 12        Analytics Engineering Manager  2023       399880.0
114               Data Science Tech Lead  2022       375000.0
130             Director of Data Science  2020       325000.0
191      Managing Director Data Science  2020       300000.0
...                                  ...   ...            ...
207                 Product Data Analyst  2020        20000.0
178  Machine Learning Research Engineer  2021        20000.0
 70                  Data Analytics Lead  2023        17511.0
160                          ML Engineer  2020        15966.0
219                  Staff Data Analyst  2020        15000.0
```

224 rows × 3 columns

In [26]:
```
list1 = ['Data Engineer', 'Data Scientist', 'Data Analyst', 'Machine Learning

salary1 = salary[salary['Job Title'].isin(list1)]
```

In [27]:

```
salary1.sort_values(by='Job Title')
```

Out[27]:

	Job Title	Year	Salary in USD
62	Data Analyst	2020	60911.166667
63	Data Analyst	2021	78258.500000
64	Data Analyst	2022	104739.781457
65	Data Analyst	2023	115299.039007
79	Data Engineer	2020	85301.384615
80	Data Engineer	2021	91636.971429
81	Data Engineer	2022	137205.879668
82	Data Engineer	2023	150907.871671
118	Data Scientist	2023	160877.946176
117	Data Scientist	2022	127960.067568
116	Data Scientist	2021	79366.230769
115	Data Scientist	2020	85970.523810
168	Machine Learning Engineer	2020	145904.500000
169	Machine Learning Engineer	2021	74611.222222
170	Machine Learning Engineer	2022	140232.355263
171	Machine Learning Engineer	2023	186091.955446
213	Research Scientist	2020	246000.000000
214	Research Scientist	2021	83003.600000
215	Research Scientist	2022	142188.733333
216	Research Scientist	2023	186193.441558

In [28]:

```
dat_anal = salary1[salary1['Job Title'] == 'Data Analyst']
dat_anal = dat_anal.reset_index()
dat_anal = dat_anal.drop('index',axis=1)
```

In [29]:

```
dat_eng = salary1[salary1['Job Title'] == 'Data Engineer']
dat_eng = dat_eng.reset_index()
dat_eng = dat_eng.drop('index',axis=1)
```

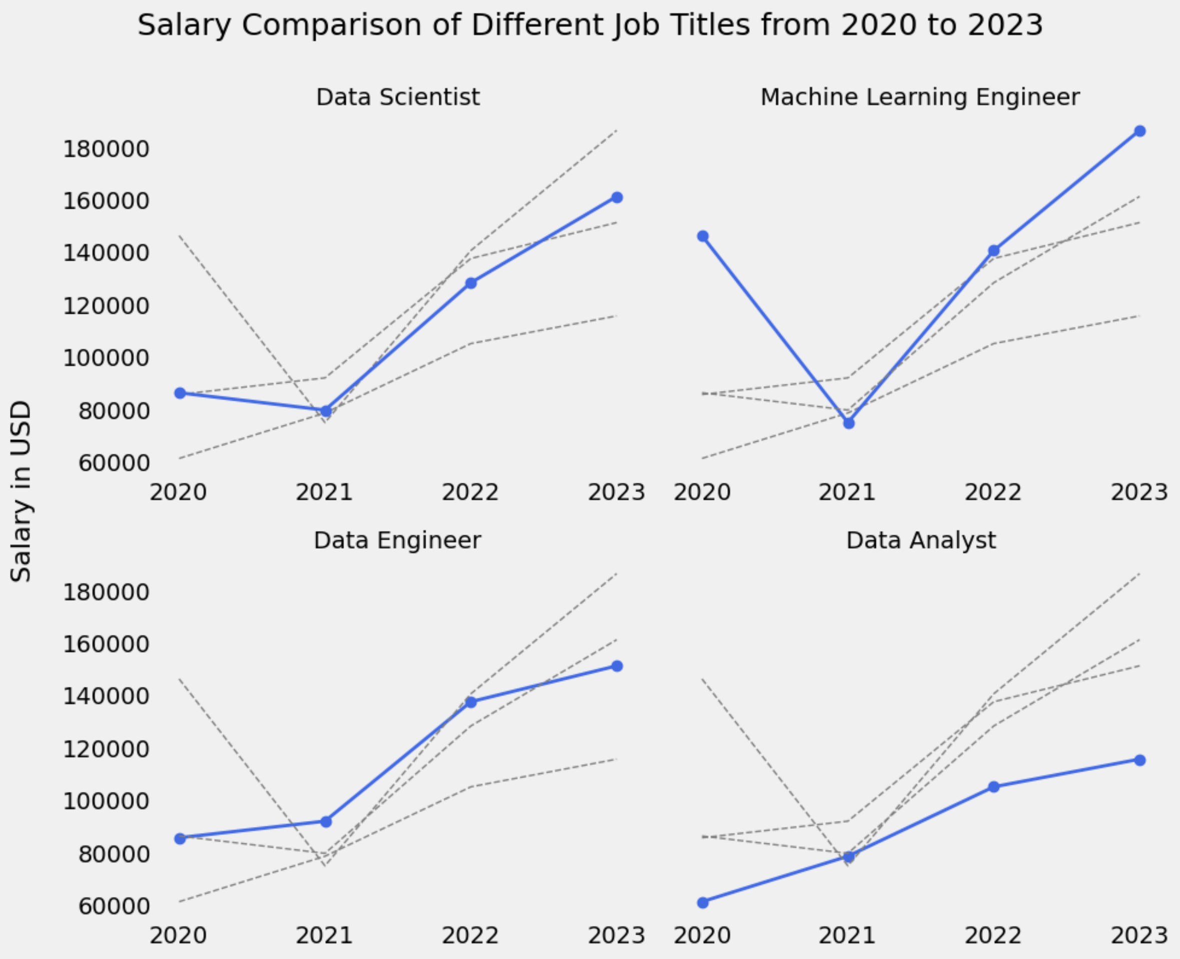
In [30]:

```
ml_eng = salary1[salary1['Job Title'] == 'Machine Learning Engineer']
ml_eng = ml_eng.reset_index()
ml_eng = ml_eng.drop('index',axis=1)
```

In [31]:

```
dat_sci = salary1[salary1['Job Title'] == 'Data Scientist']
dat_sci = dat_sci.reset_index()
dat_sci = dat_sci.drop('index',axis=1)
```

```
In [35]: years = [2020, 2021, 2022, 2023]
         plt.style.use('fivethirtyeight')
         fig, ax = plt.subplots(2, 2, sharey=True, figsize=(10,8), )
         plt.suptitle('Salary Comparison of Different Job Titles from 2020 to 2023', fo
         plt.grid(visible=None)

         ax[0,0].plot('Year','Salary in USD', data = dat_anal, linestyle = 'dashed', c=
         ax[0,0].plot('Year','Salary in USD', data = dat_eng, linestyle = 'dashed', c='
         ax[0,0].plot('Year','Salary in USD', data = ml_eng, linestyle = 'dashed', c='g
         ax[0,0].plot('Year','Salary in USD', data = dat_sci, linestyle = 'solid', c='r
         ax[0,0].set_title('Data Scientist', fontsize=14, color='black')
         ax[0,0].set_xticks(years)
         ax[0,0].grid(False)

         ax[0,1].plot('Year','Salary in USD', data = dat_anal, linestyle = 'dashed', c=
         ax[0,1].plot('Year','Salary in USD', data = dat_eng, linestyle = 'dashed', c='
         ax[0,1].plot('Year','Salary in USD', data = ml_eng, linestyle = 'solid', c='ro
         ax[0,1].plot('Year','Salary in USD', data = dat_sci, linestyle = 'dashed', c='
         ax[0,1].set_title('Machine Learning Engineer', fontsize=14, color='black')
         ax[0,1].set_xticks(years)
         ax[0,1].grid(False)


         ax[1,0].plot('Year','Salary in USD', data = dat_anal, linestyle = 'dashed', c=
         ax[1,0].plot('Year','Salary in USD', data = dat_eng, linestyle = 'solid', c='r
         ax[1,0].plot('Year','Salary in USD', data = ml_eng, linestyle = 'dashed', c='g
         ax[1,0].plot('Year','Salary in USD', data = dat_sci, linestyle = 'dashed', c='
         ax[1,0].set_title('Data Engineer', fontsize=14, color='black')
         ax[1,0].set_xticks(years)
         ax[1,0].grid(False)

         ax[1,1].plot('Year','Salary in USD', data = dat_anal, linestyle = 'solid', c='
         ax[1,1].plot('Year','Salary in USD', data = dat_eng, linestyle = 'dashed', c='
         ax[1,1].plot('Year','Salary in USD', data = ml_eng, linestyle = 'dashed', c='g
         ax[1,1].plot('Year','Salary in USD', data = dat_sci, linestyle = 'dashed', c='
         ax[1,1].set_title('Data Analyst', fontsize=14, color='black')
         ax[1,1].set_xticks(years)
         ax[1,1].grid(False)



         fig.supylabel('Salary in USD', fontsize=16, color='black')
         fig.tight_layout()
         plt.show

Out[35]: <function matplotlib.pyplot.show(close=None, block=None)>
```

Salary Comparison of Different Job Titles from 2020 to 2023
Data Scientist
Machine Learning Engineer
Data Engineer
Data Analyst
Salary in USD
180000
160000
140000
120000
100000
80000
60000
2020
2021
2022
2023