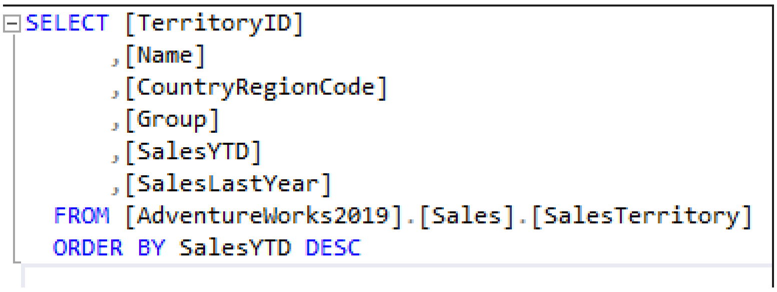
# Imports

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

# Q1

***Question 1:*** What are the regional sales in the best performing country?

**SQL:** Selecting required data from corresponding table(s)



```
SELECT [TerritoryID]
      ,[Name]
      ,[CountryRegionCode]
      ,[Group]
      ,[SalesYTD]
      ,[SalesLastYear]
  FROM [AdventureWorks2019].[Sales].[SalesTerritory]
  ORDER BY SalesYTD DESC
```

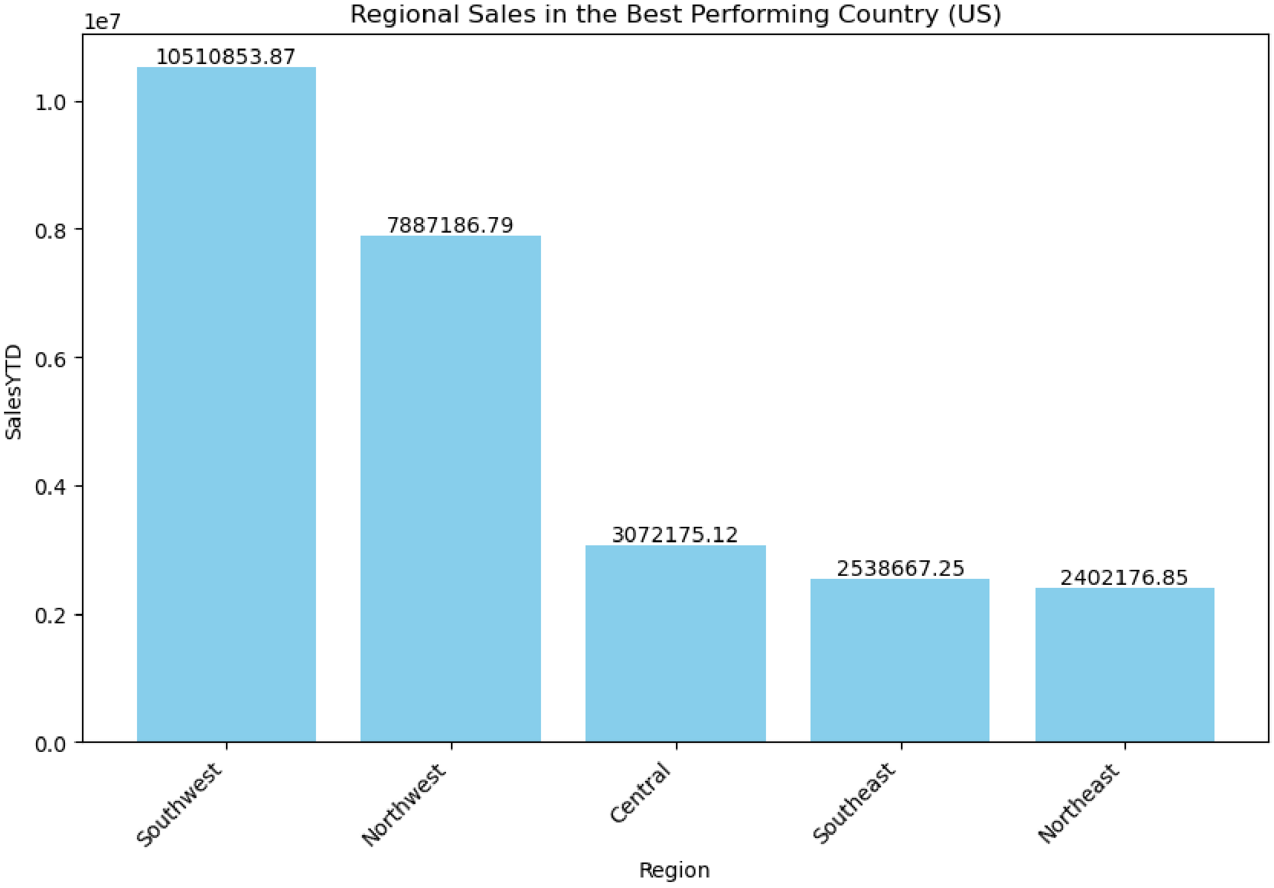
```
q1_data = pd.read_csv("C:/Users/amoha/OneDrive/Desktop/Gen_Project/q1_data.csv")
q1_data.head(2)
```

| | TerritoryID | Name | CountryRegionCode | Group | SalesYTD | SalesLastYear |
|---|---|---|---|---|---|---|
| **0** | 4 | Southwest | US | North America | 1.051085e+07 | 5.366576e+06 |
| **1** | 1 | Northwest | US | North | 7.887187e+06 | 3.298694e+06 |

```
q1_data.groupby('CountryRegionCode')['SalesYTD'].sum()
```

```
CountryRegionCode
AU    5.977815e+06
CA    6.771829e+06
DE    3.805202e+06
FR    4.772398e+06
GB    5.012905e+06
US    2.641106e+07
Name: SalesYTD, dtype: float64
```

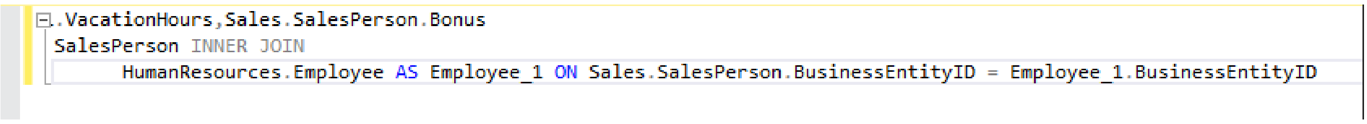
```
best_performing_country = q1_data.groupby('CountryRegionCode')['SalesYTD'].sum().idxm
best_performing_country_data = q1_data[q1_data['CountryRegionCode'] == best_performin


plt.figure(figsize=(10, 6))
bars = plt.bar(best_performing_country_data['Name'], best_performing_country_data['Sa
plt.title(f'Regional Sales in the Best Performing Country ({best_performing_country})
plt.xlabel('Region')
plt.ylabel('SalesYTD')
plt.xticks(rotation=45, ha='right')

for bar in bars:
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width()/2, yval, round(yval, 2), ha='center', va='

plt.show()
```



## Q2

***Question 2:*** What is the relationship between annual leave taken and bonus?

**SQL:** Selecting required data from corresponding table(s)



```
⊟.VacationHours,Sales.SalesPerson.Bonus
 SalesPerson INNER JOIN
        HumanResources.Employee AS Employee_1 ON Sales.SalesPerson.BusinessEntityID = Employee_1.BusinessEntityID
```

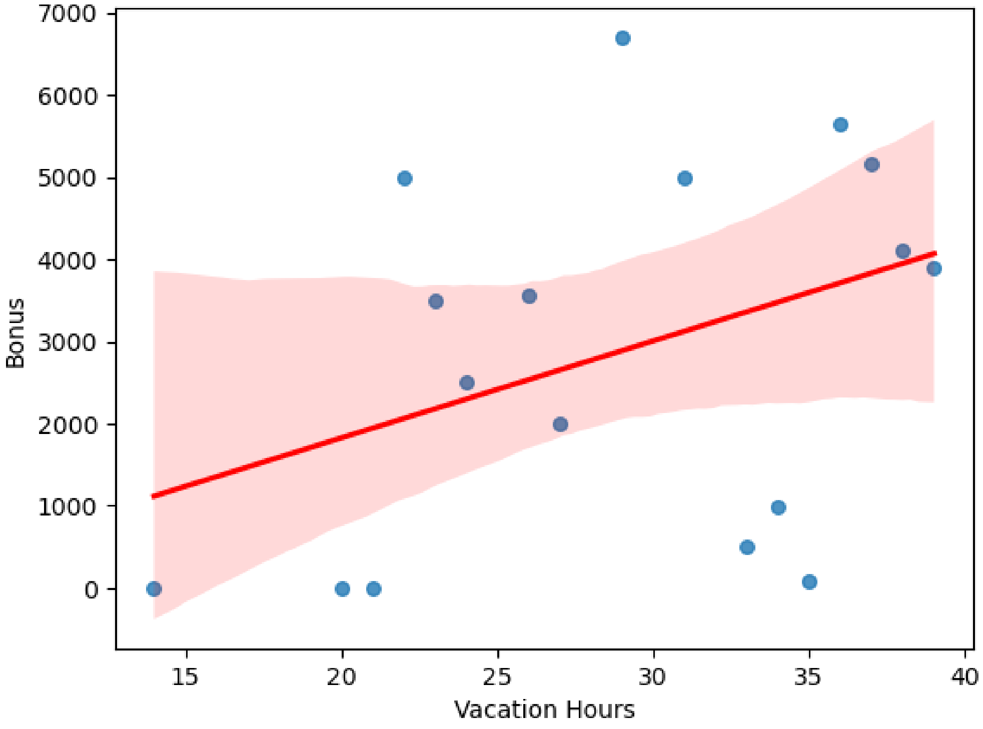
```
q2_data = pd.read_csv("C:/Users/amoha/OneDrive/Desktop/Gen_Project/q2_data.csv")
q2_data.head(2)
```

|   | VacationHours | Bonus |
|---|---|---|
| **0** | 14 | 0 |
| **1** | 38 | 4100 |

```
correlation = q2_data['VacationHours'].corr(q2_data['Bonus'])
print(f'Correlation between Vacation Hours and Bonus: {correlation}')

sns.regplot(x='VacationHours', y='Bonus', data=q2_data, scatter_kws={'s': 30}, line_k
plt.title(f'Relationship between Vacation Hours and Bonus (Correlation  = {correlatio
plt.xlabel('Vacation Hours')
plt.ylabel('Bonus')
plt.show()
```

```
Correlation between Vacation Hours and Bonus: 0.3821074616559863
```

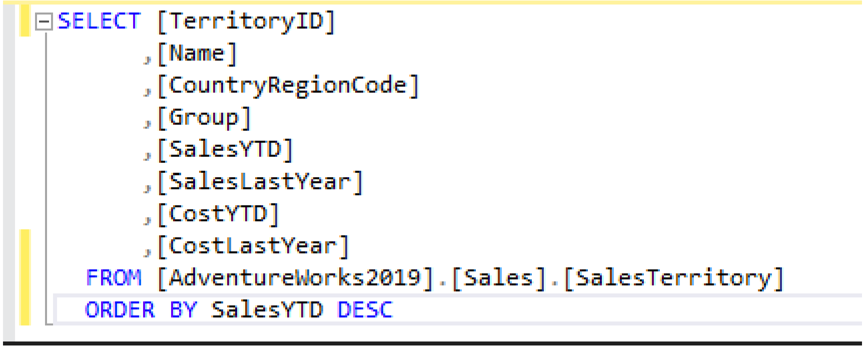
Relationship between Vacation Hours and Bonus (Correlation = 0.3821074616559863)



# Q3

***Q3 :*** What is the relationship between Country and Revenue?

**SQL:** Selecting required data from corresponding table(s)

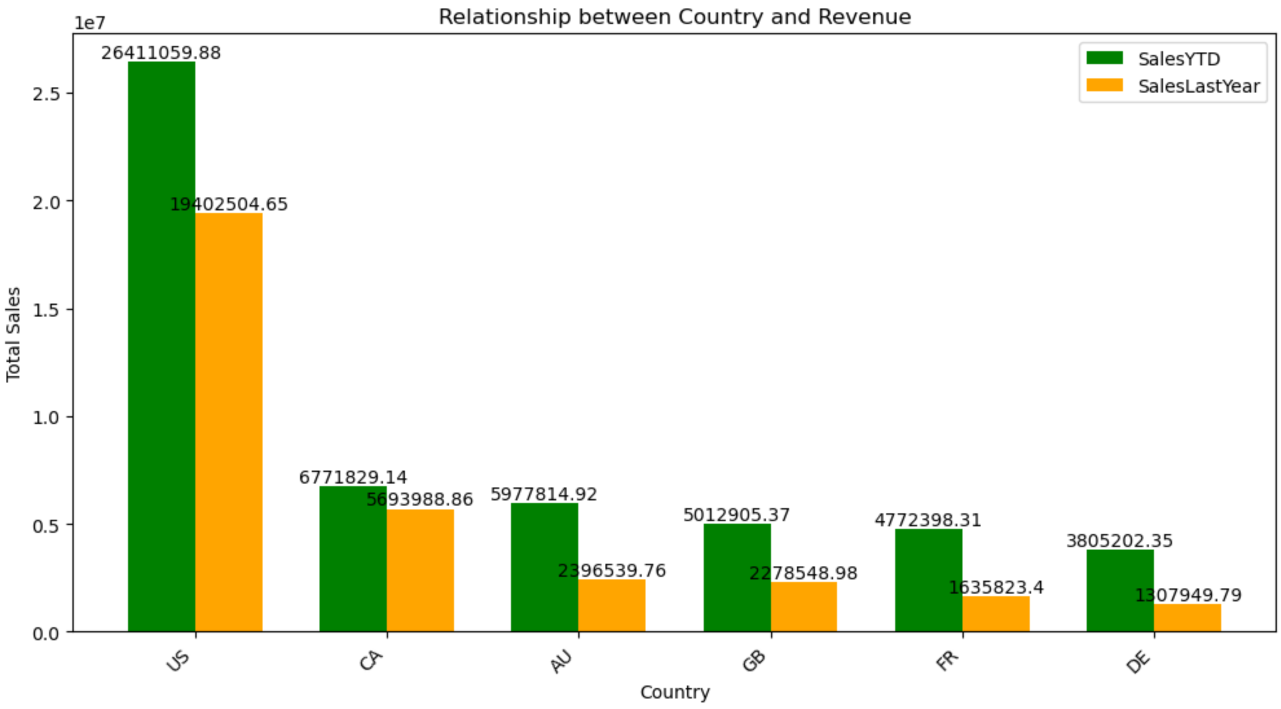


```
SELECT [TerritoryID]
      ,[Name]
      ,[CountryRegionCode]
      ,[Group]
      ,[SalesYTD]
      ,[SalesLastYear]
      ,[CostYTD]
      ,[CostLastYear]
  FROM [AdventureWorks2019].[Sales].[SalesTerritory]
  ORDER BY SalesYTD DESC
```

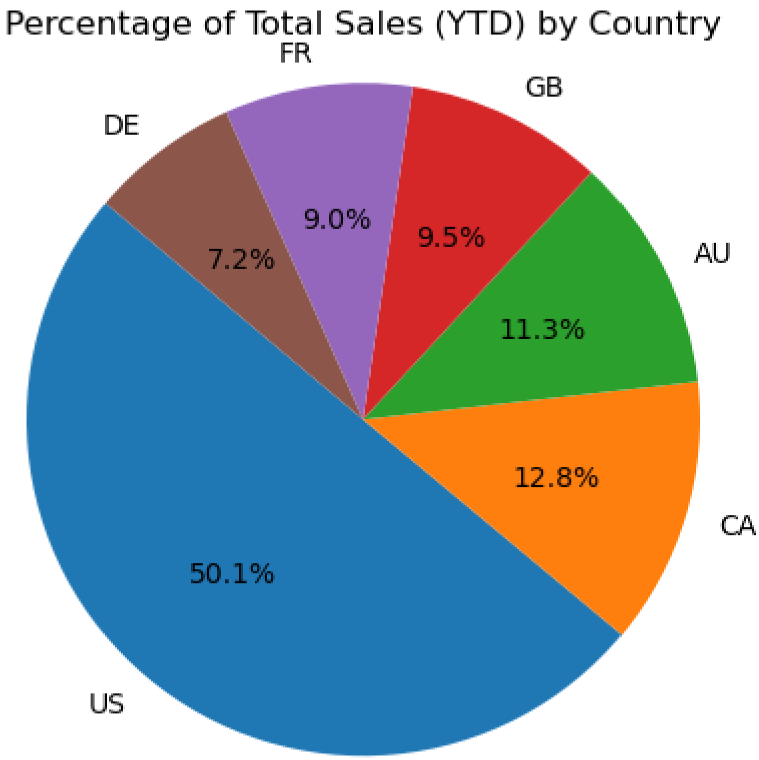
```
q3_data = pd.read_csv("C:/Users/amoha/OneDrive/Desktop/Gen_Project/q3_data.csv")
q3_data.head(2)
```

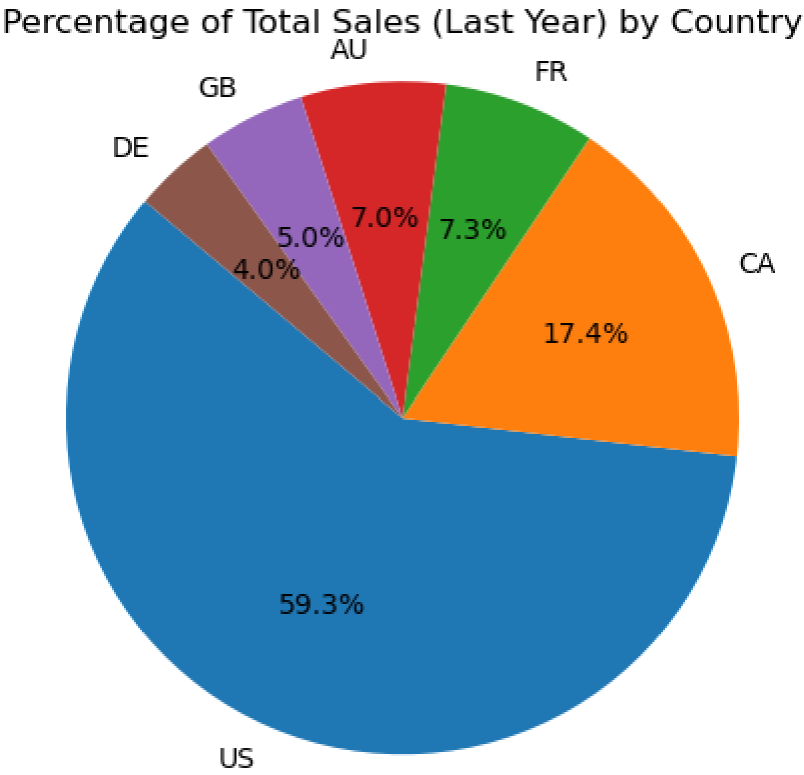
| | TerritoryID | Name | CountryRegionCode | Group | SalesYTD | SalesLastYear |
|---|---|---|---|---|---|---|
| **0** | 4 | Southwest | US | North America | 1.051085e+07 | 5.366576e+06 |

```
total_rev = q3_data.groupby('CountryRegionCode')['SalesYTD'].sum()
total_rev = total_rev.sort_values(ascending=False)

total_rev_ly = q3_data.groupby('CountryRegionCode')['SalesLastYear'].sum()
total_rev_ly = total_rev_ly.sort_values(ascending=False)
bar_width = 0.35

indices = np.arange(len(total_rev))

plt.figure(figsize=(12, 6))
plt.title(f'Relationship between Country and Revenue')
plt.xlabel('Country')
plt.ylabel('Total Sales')

plt.bar(indices, total_rev, width=bar_width, color='green', label='SalesYTD')
plt.bar(indices + bar_width, total_rev_ly, width=bar_width, color='orange', label='Sa

for idx, value in enumerate(total_rev):
    plt.text(idx, value, round(value, 2), ha='center', va='bottom')

for idx, value in enumerate(total_rev_ly):
    plt.text(idx + bar_width, value, round(value, 2), ha='center', va='bottom')

plt.xticks(indices + bar_width / 2, total_rev.index, rotation=45, ha='right')
plt.legend()
plt.show()
```



```
plt.title('Percentage of Total Sales (YTD) by Country')
plt.pie(total_rev, labels=total_rev.index, autopct='%1.1f%%', startangle=140)
plt.axis('equal')
plt.show()


plt.title('Percentage of Total Sales (Last Year) by Country')
plt.pie(total_rev_ly, labels=total_rev_ly.index, autopct='%1.1f%%', startangle=140)
plt.axis('equal')
plt.show()
```

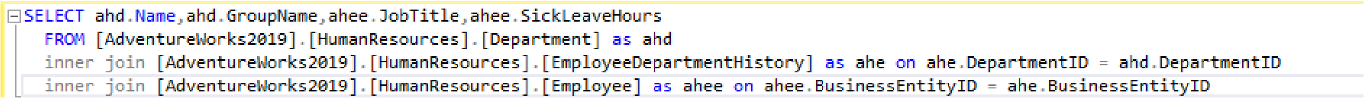




## Q4

***Q4 :*** What is the relationship between sick leave and Job Title (PersonType)?

**SQL:** Selecting required data from corresponding table(s)



```
SELECT ahd.Name,ahd.GroupName,ahee.JobTitle,ahee.SickLeaveHours
  FROM [AdventureWorks2019].[HumanResources].[Department] as ahd
  inner join [AdventureWorks2019].[HumanResources].[EmployeeDepartmentHistory] as ahe on ahe.DepartmentID = ahd.DepartmentID
  inner join [AdventureWorks2019].[HumanResources].[Employee] as ahee on ahee.BusinessEntityID = ahe.BusinessEntityID
```

```
q4_data = pd.read_csv("C:/Users/amoha/OneDrive/Desktop/Generation/Gen_Project/project
print(len(q4_data.JobTitle.unique()))
q4_data.head(2)
```

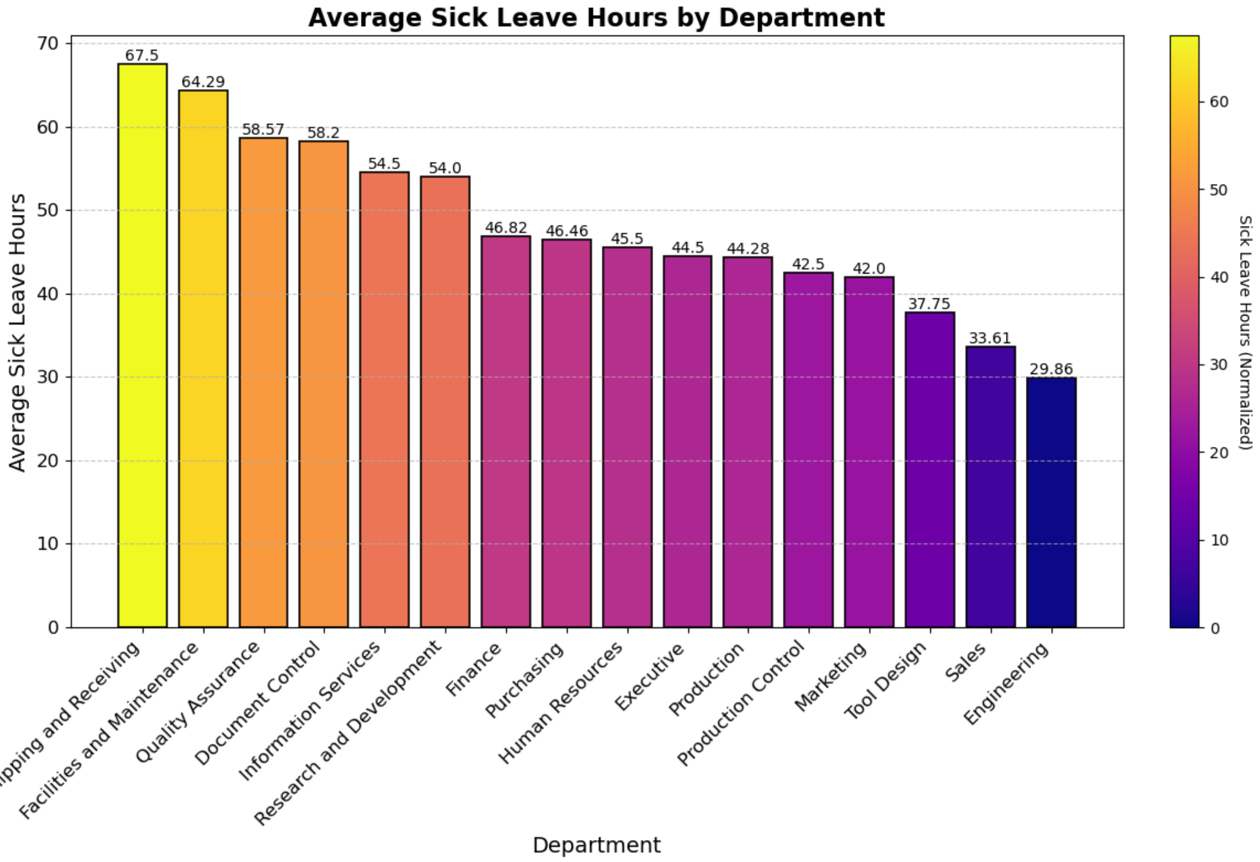
```
67
```

| | Name | GroupName | JobTitle | SickLeaveHours |
|---|---|---|---|---|
| **0** | Engineering | Research and Development | Vice President of Engineering | 20 |
| **1** | Engineering | Research and Development | Engineering Manager | 21 |

```
q4_grouped = pd.DataFrame(q4_data.groupby("Name")["SickLeaveHours"].mean())
```

```
q4_grouped = q4_grouped.sort_values("SickLeaveHours",axis = 0,ascending= False)
```

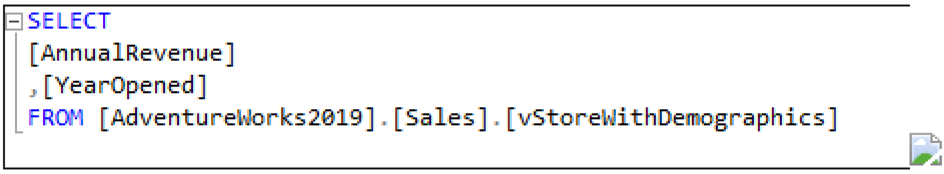
```
import seaborn as sns
import matplotlib.pyplot as plt
from matplotlib import cm
from matplotlib.colors import Normalize


colors = plt.cm.plasma(Normalize()(q4_grouped.SickLeaveHours.values))

plt.figure(figsize=(12, 8))
bars = plt.bar(q4_grouped.SickLeaveHours.index, q4_grouped.SickLeaveHours.values, col
plt.title("Average Sick Leave Hours by Department", fontsize=16, fontweight='bold')
plt.xlabel("Department", fontsize=14)
plt.ylabel("Average Sick Leave Hours", fontsize=14)
plt.xticks(rotation=45, ha='right', fontsize=12)
plt.yticks(fontsize=12)
plt.grid(axis='y', linestyle='--', alpha=0.7)

for bar, value, color in zip(bars, q4_grouped.SickLeaveHours, colors):
    plt.text(bar.get_x() + bar.get_width() / 2, value, f"{round(value, 2)}", ha='cent


sm = plt.cm.ScalarMappable(cmap='plasma', norm=Normalize(vmin=0, vmax=max(q4_grouped.
sm.set_array([])
cbar = plt.colorbar(sm, ax=plt.gca(), fraction=0.046, pad=0.04)
cbar.set_label('Sick Leave Hours (Normalized)', rotation=270, labelpad=15)

plt.tight_layout()
plt.show()
```



## Q5

***Q5 :*** What is the relationship between store trading duration and revenue?

**SQL:** Selecting required data from corresponding table(s)



```sql
SELECT
[AnnualRevenue]
,[YearOpened]
FROM [AdventureWorks2019].[Sales].[vStoreWithDemographics]
```

image.png

```python
q5_data = pd.read_csv("C:/Users/amoha/OneDrive/Desktop/Gen_Project/q5_data.csv")
q5_data["Years"] = 2023 - q5_data["YearOpened"]
q5_data.head()
```

| | Name | AnnualRevenue | YearOpened | Years |
|---|---|---|---|---|
| **0** | Purchase Mart | 150000.0 | 1992 | 31 |
| **1** | Major Sport Suppliers | 300000.0 | 1998 | 25 |
| **2** | Family's Favorite Bike Shop | 80000.0 | 1997 | 26 |
| **3** | Global Plaza | 80000.0 | 1975 | 48 |
| **4** | Imported and Domestic Cycles | 100000.0 | 2000 | 23 |

```python
q5_data[q5_data.Name.duplicated()]
```

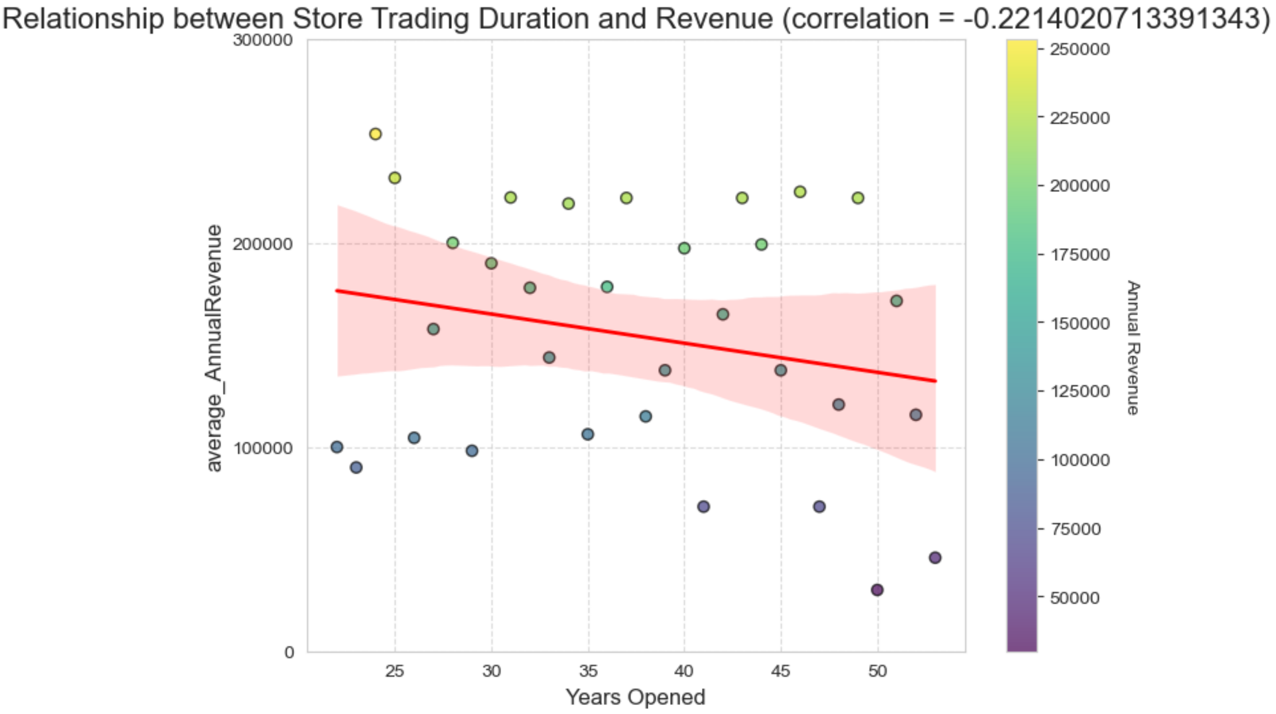
| | Name | AnnualRevenue | YearOpened | Years |
|---|---|---|---|---|
| **540** | Friendly Bike Shop | 150000.0 | 1996 | 27 |
| **558** | Sports Products Store | 300000.0 | 1999 | 24 |

```python
q5_data[q5_data.Name == "Friendly Bike Shop"]
```

| | Name | AnnualRevenue | YearOpened | Years |
|---|---|---|---|---|
| **235** | Friendly Bike Shop | 300000.0 | 1980 | 43 |
| **540** | Friendly Bike Shop | 150000.0 | 1996 | 27 |

```python
q5_data_grouped = q5_data.groupby("Years")["AnnualRevenue"].mean().reset_index()
q5_data_grouped.columns = ["Years", "average_AnnualRevenue"]
```

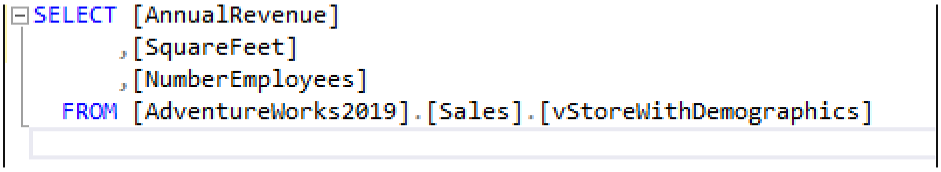
```
sns.set_style("whitegrid")

fig, ax = plt.subplots(figsize=(8, 6))

scatter = ax.scatter(q5_data_grouped['Years'], q5_data_grouped['average_AnnualRevenue
correlation = q5_data_grouped['Years'].corr(q5_data_grouped['average_AnnualRevenue'])

cbar = plt.colorbar(scatter)
cbar.set_label('Annual Revenue', rotation=270, labelpad=15)

sns.regplot(x='Years', y='average_AnnualRevenue', data=q5_data_grouped, scatter=False

ax.set_title(f"Relationship between Store Trading Duration and Revenue (correlation =
ax.set_xlabel("Years Opened", fontsize=12)
ax.set_ylabel("average_AnnualRevenue", fontsize=12)
ax.set_yticks([0, 100000, 200000, 300000])
ax.grid(True, linestyle='--', alpha=0.7)
ax.tick_params(axis='both', which='major', labelsize=10)

plt.show()
```



# Q6

***Q6 :*** What is the relationship between the size of the stores, number of employees and revenue?

**SQL:** Selecting required data from corresponding table(s)



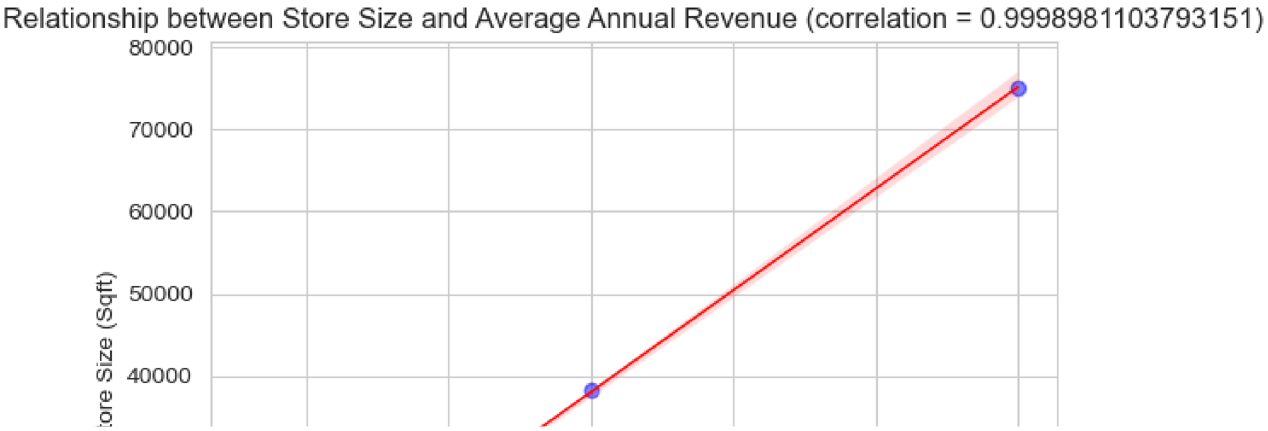
```
SELECT [AnnualRevenue]
      ,[SquareFeet]
      ,[NumberEmployees]
  FROM [AdventureWorks2019].[Sales].[vStoreWithDemographics]
```

```
q6_data = pd.read_csv("C:/Users/amoha/OneDrive/Desktop/Gen_Project/q6_data.csv")
q6_data.head()
```

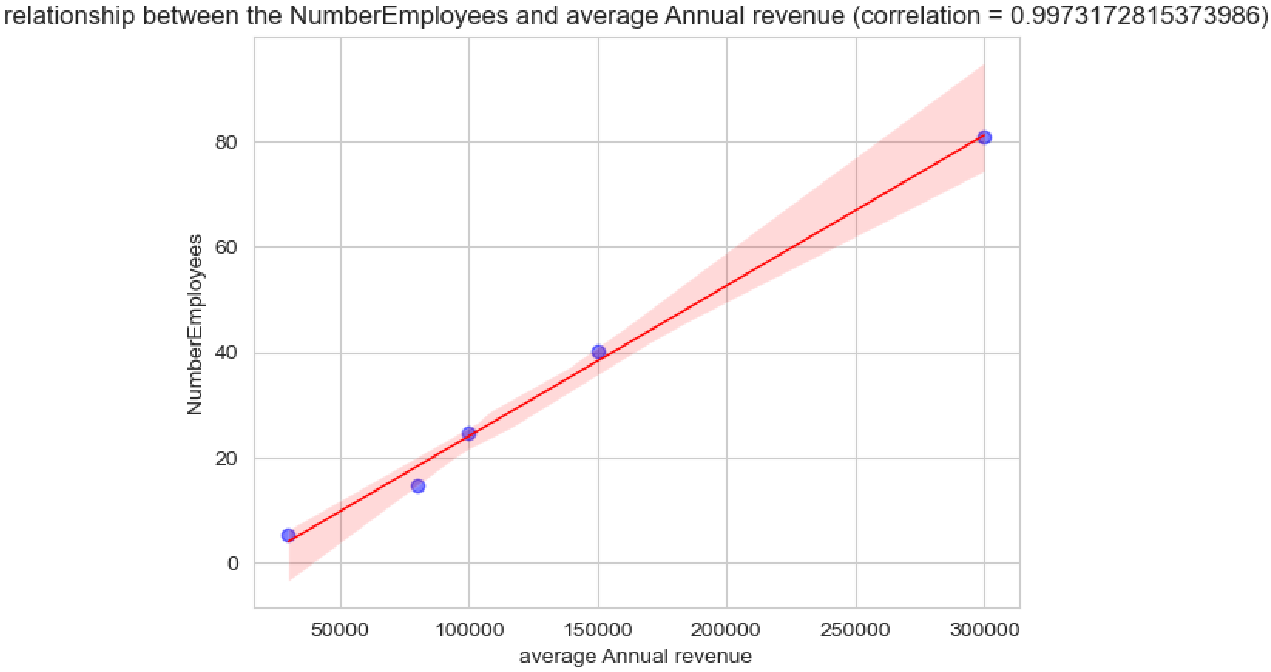
| | AnnualRevenue | SquareFeet | NumberEmployees |
|---|---|---|---|
| **0** | 80000.0 | 21000 | 13 |
| **1** | 80000.0 | 18000 | 14 |
| **2** | 80000.0 | 21000 | 15 |
| **3** | 80000.0 | 18000 | 16 |
| **4** | 80000.0 | 21000 | 17 |

```
q6_SquareFeet_grouped = q6_data.groupby("AnnualRevenue")["SquareFeet"].mean().reset_i
q6_SquareFeet_grouped.columns = ["average_AnnualRevenue", "SquareFeet"]
```

```
plt.scatter(q6_SquareFeet_grouped['average_AnnualRevenue'], q6_SquareFeet_grouped['Sq
sns.regplot(x='average_AnnualRevenue', y='SquareFeet', data=q6_SquareFeet_grouped, sc
correlation = q6_SquareFeet_grouped['average_AnnualRevenue'].corr(q6_SquareFeet_group
plt.title(f"Relationship between Store Size and Average Annual Revenue (correlation =
plt.xlabel("Average Annual Revenue")
plt.ylabel("Store Size (Sqft)")

plt.show()
```



```
q6_Employees_grouped = q6_data.groupby("AnnualRevenue")["NumberEmployees"].mean().res
q6_Employees_grouped.columns = ["average_AnnualRevenue", "Employees"]

plt.scatter(q6_Employees_grouped.average_AnnualRevenue,q6_Employees_grouped.Employees
sns.regplot(x='average_AnnualRevenue', y='Employees', data=q6_Employees_grouped, scat
correlation = q6_Employees_grouped.average_AnnualRevenue.corr(q6_Employees_grouped.Em

plt.title(f"relationship between the NumberEmployees and average Annual revenue (corr
plt.xlabel("average Annual revenue")
plt.ylabel("NumberEmployees")

plt.show()
```


relationship between the NumberEmployees and average Annual revenue (correlation = 0.9973172815373986)
NumberEmployees
average Annual revenue