Name : Tavrez Alam Ansari

Roll no : 61

Branch : IT

Semester : 4

Subject : Tools For Data Science

In [3]:
```
import pandas as pd
import matplotlib.pyplot as plt
```

In [4]:
```
data=pd.read_excel("C:\\Users\\SHAHNAWAZ\\downloads\\new.xlsx")
```

In [23]:
```
data
```

Out[23]:

	Segment	Country	Product	Discount Band	Units Sold	Manufacturing Price	Sale Price	Gross Sales	Discounts	Sales	COGS	Profit	Date	Month Number	Month Name	Year
0	Government	Canada	Carretera	NaN	1618.5	3	20	32370.0	0.00	32370.00	16185.0	16185.00	2014-01-01	1	January	2014
1	Government	Germany	Carretera	NaN	1321.0	3	20	26420.0	0.00	26420.00	13210.0	13210.00	2014-01-01	1	January	2014
2	Midmarket	France	Carretera	NaN	2178.0	3	15	32670.0	0.00	32670.00	21780.0	10890.00	2014-06-01	6	June	2014
3	Midmarket	Germany	Carretera	NaN	888.0	3	15	13320.0	0.00	13320.00	8880.0	4440.00	2014-06-01	6	June	2014
4	Midmarket	Mexico	Carretera	NaN	2470.0	3	15	37050.0	0.00	37050.00	24700.0	12350.00	2014-06-01	6	June	2014
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
695	Small Business	France	Amarilla	High	2475.0	260	300	742500.0	111375.00	631125.00	618750.0	12375.00	2014-03-01	3	March	2014
696	Small Business	Mexico	Amarilla	High	546.0	260	300	163800.0	24570.00	139230.00	136500.0	2730.00	2014-10-01	10	October	2014
697	Government	Mexico	Montana	High	1368.0	5	7	9576.0	1436.40	8139.60	6840.0	1299.60	2014-02-01	2	February	2014
698	Government	Canada	Paseo	High	723.0	10	7	5061.0	759.15	4301.85	3615.0	686.85	2014-04-01	4	April	2014
699	Channel Partners	United States of America	VTT	High	1806.0	250	12	21672.0	3250.80	18421.20	5418.0	13003.20	2014-05-01	5	May	2014

700 rows × 16 columns

In [24]:
```
data.head()
```

Out[24]:

	Segment	Country	Product	Discount Band	Units Sold	Manufacturing Price	Sale Price	Gross Sales	Discounts	Sales	COGS	Profit	Date	Month Number	Month Name	Year
0	Government	Canada	Carretera	NaN	1618.5	3	20	32370.0	0.0	32370.0	16185.0	16185.0	2014-01-01	1	January	2014
1	Government	Germany	Carretera	NaN	1321.0	3	20	26420.0	0.0	26420.0	13210.0	13210.0	2014-01-01	1	January	2014
2	Midmarket	France	Carretera	NaN	2178.0	3	15	32670.0	0.0	32670.0	21780.0	10890.0	2014-06-01	6	June	2014
3	Midmarket	Germany	Carretera	NaN	888.0	3	15	13320.0	0.0	13320.0	8880.0	4440.0	2014-06-01	6	June	2014
4	Midmarket	Mexico	Carretera	NaN	2470.0	3	15	37050.0	0.0	37050.0	24700.0	12350.0	2014-06-01	6	June	2014

In [25]:
```
data.tail()
```

Out[25]:

	Segment	Country	Product	Discount Band	Units Sold	Manufacturing Price	Sale Price	Gross Sales	Discounts	Sales	COGS	Profit	Date	Month Number	Month Name	Year
695	Small Business	France	Amarilla	High	2475.0	260	300	742500.0	111375.00	631125.00	618750.0	12375.00	2014-03-01	3	March	2014
696	Small Business	Mexico	Amarilla	High	546.0	260	300	163800.0	24570.00	139230.00	136500.0	2730.00	2014-10-01	10	October	2014
697	Government	Mexico	Montana	High	1368.0	5	7	9576.0	1436.40	8139.60	6840.0	1299.60	2014-02-01	2	February	2014
698	Government	Canada	Paseo	High	723.0	10	7	5061.0	759.15	4301.85	3615.0	686.85	2014-04-01	4	April	2014
699	Channel Partners	United States of America	VTT	High	1806.0	250	12	21672.0	3250.80	18421.20	5418.0	13003.20	2014-05-01	5	May	2014

In [12]:
```

```

Out[12]:

	Username; Identifier;One-time password;Recovery code;First name;Last name;Department;Location
0	booker12;9012;12se74;rb9012;Rachel;Booker;Sale...
1	grey07;2070;04ap67;lg2070;Laura;Grey;Depot;London
2	johnson81;4081;30no86;cj4081;Craig;Johnson;Dep...
3	jenkins46;9346;14ju73;mj9346;Mary;Jenkins;Engi...
4	smith79;5079;09ja61;js5079;Jamie;Smith;Enginee...

In [26]:
```
data.isnull().sum()
```

Out[26]:
```
Segment                0
Country                0
Product                0
Discount Band         53
Units Sold             0
Manufacturing Price    0
Sale Price             0
Gross Sales            0
Discounts              0
 Sales                 0
COGS                   0
Profit                 0
Date                   0
Month Number           0
Month Name             0
Year                   0
dtype: int64
```

In [27]:
```
data.fillna(0)
```

Out[27]:

	Segment	Country	Product	Discount Band	Units Sold	Manufacturing Price	Sale Price	Gross Sales	Discounts	Sales	COGS	Profit	Date	Month Number	Month Name	Year
0	Government	Canada	Carretera	0	1618.5	3	20	32370.0	0.00	32370.00	16185.0	16185.00	2014-01-01	1	January	2014
1	Government	Germany	Carretera	0	1321.0	3	20	26420.0	0.00	26420.00	13210.0	13210.00	2014-01-01	1	January	2014
2	Midmarket	France	Carretera	0	2178.0	3	15	32670.0	0.00	32670.00	21780.0	10890.00	2014-06-01	6	June	2014
3	Midmarket	Germany	Carretera	0	888.0	3	15	13320.0	0.00	13320.00	8880.0	4440.00	2014-06-01	6	June	2014
4	Midmarket	Mexico	Carretera	0	2470.0	3	15	37050.0	0.00	37050.00	24700.0	12350.00	2014-06-01	6	June	2014
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
695	Small Business	France	Amarilla	High	2475.0	260	300	742500.0	111375.00	631125.00	618750.0	12375.00	2014-03-01	3	March	2014
696	Small Business	Mexico	Amarilla	High	546.0	260	300	163800.0	24570.00	139230.00	136500.0	2730.00	2014-10-01	10	October	2014
697	Government	Mexico	Montana	High	1368.0	5	7	9576.0	1436.40	8139.60	6840.0	1299.60	2014-02-01	2	February	2014
698	Government	Canada	Paseo	High	723.0	10	7	5061.0	759.15	4301.85	3615.0	686.85	2014-04-01	4	April	2014
699	Channel Partners	United States of America	VTT	High	1806.0	250	12	21672.0	3250.80	18421.20	5418.0	13003.20	2014-05-01	5	May	2014

700 rows × 16 columns

In [30]:
```
mean=data["Units Sold"].mean()
print("mean : ",mean)
```

```
mean :  1608.2942857142857
```

In [32]:
```
median=data["Units Sold"].median()
print("Median : ",median)
```

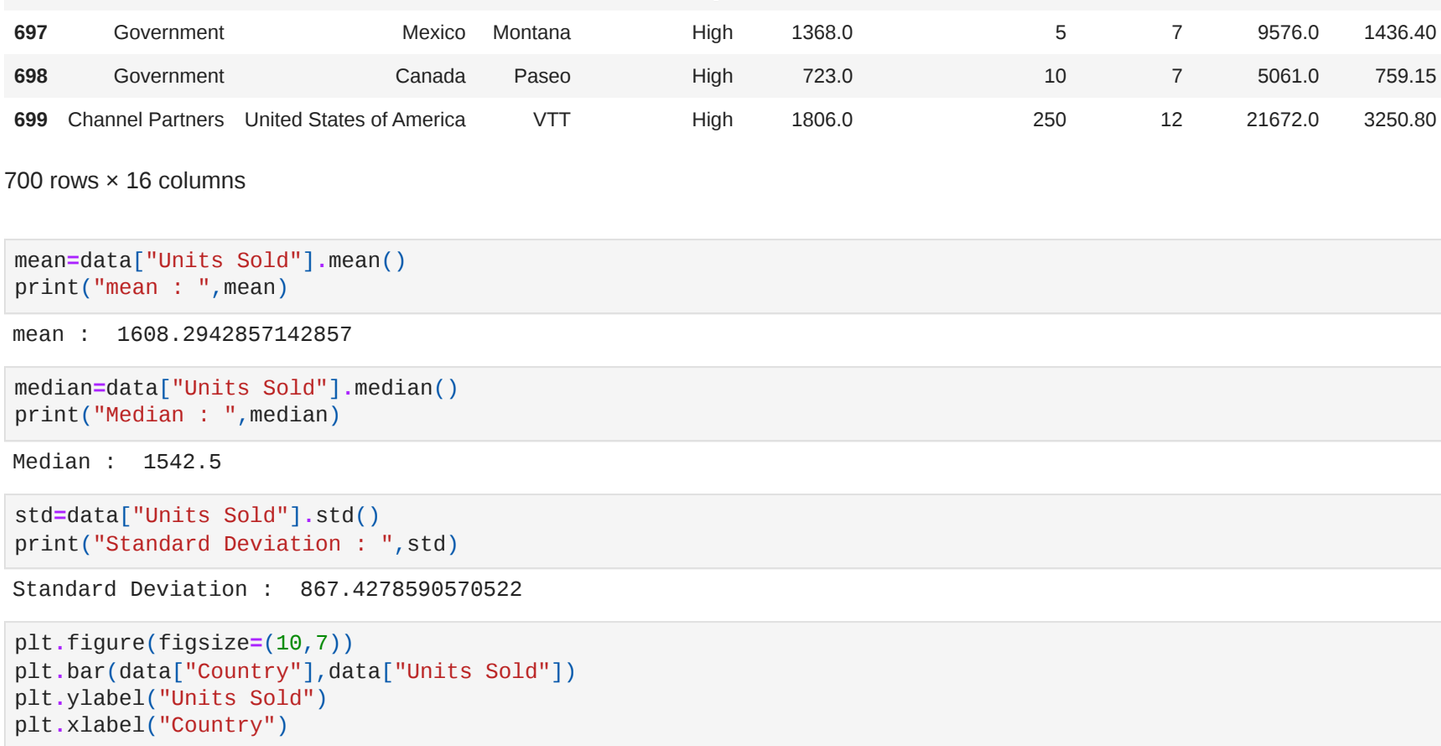
```
Median :  1542.5
```

In [33]:
```
std=data["Units Sold"].std()
print("Standard Deviation : ",std)
```

```
Standard Deviation :  867.4278590570522
```

In [5]:
```
plt.figure(figsize=(10,7))
plt.bar(data["Country"],data["Units Sold"])
plt.ylabel("Units Sold")
plt.xlabel("Country")
```

Out[5]:
```
Text(0.5, 0, 'Country')
```

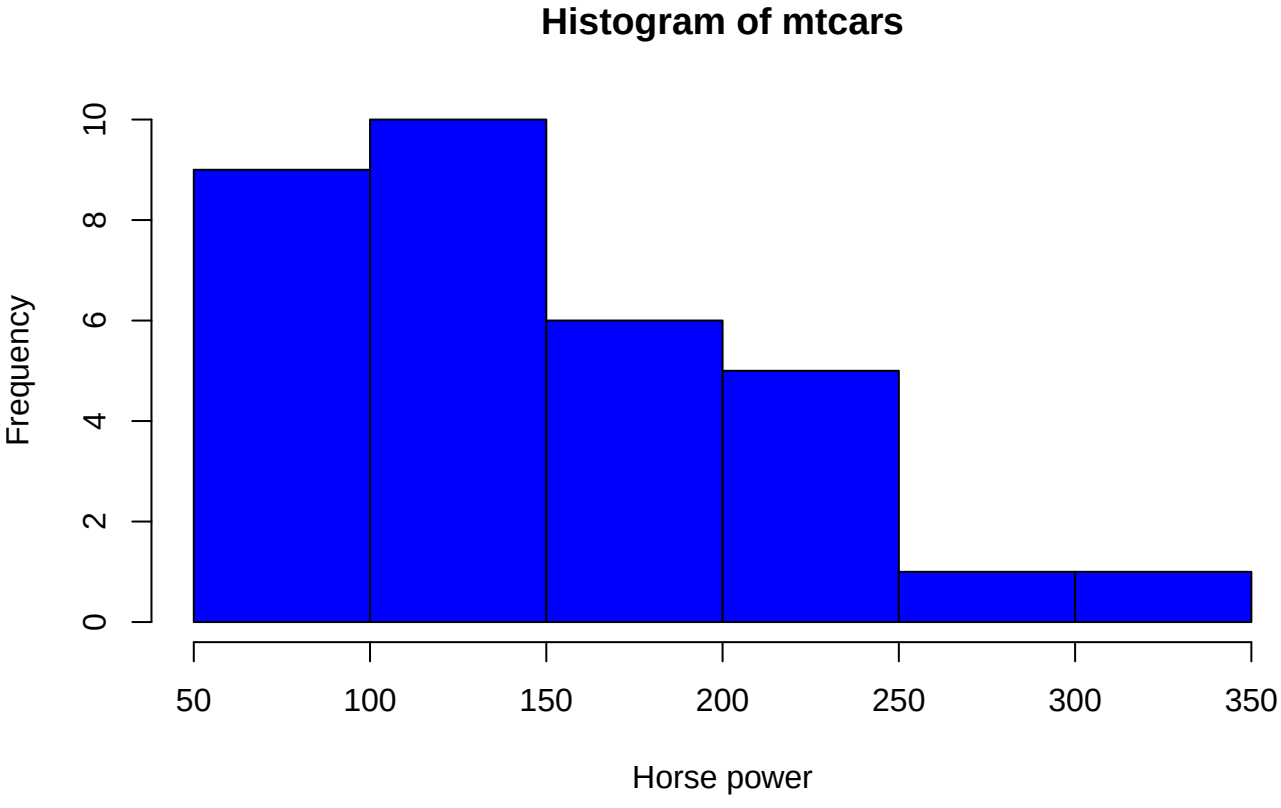


In [6]:
```
#Therefor from above diagram we can conclude that United states of America had sold maximum numbers of units
```

```
mtcars
data<-mtcars
mean1<-mean(data$wt)
paste("mean : ",mean1)

paste("meadian : ",median(data$wt))

paste("Standard Deviation",sd(data$wt))

auto<-mtcars[mtcars$am==0,]
manual <- mtcars[mtcars$am == 1, ]

t.test(mpg ~ am, data = mtcars)

hist(data$hp,xlab="Horse power",main="Histogram of
mtcars",col='blue',border='black',breaks=10)

print("By the given Hypothesis Testin we can say that The average mpg is different between
automatic and manual transmission cars.")
```

```
> mtcars
                     mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4           21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag       21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710          22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive      21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout   18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant             18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360          14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 240D           24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
Merc 230            22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
Merc 280            19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
Merc 280C           17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
Merc 450SE          16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
Merc 450SL          17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
Merc 450SLC         15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
Fiat 128            32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
Honda Civic         30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
Toyota Corolla      33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
Toyota Corona       21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
Dodge Challenger    15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
AMC Javelin         15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
Camaro Z28          13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
Pontiac Firebird    19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
Fiat X1-9           27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
Porsche 914-2       26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
Lotus Europa        30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
Ford Pantera L      15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
Ferrari Dino        19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
Maserati Bora       15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
Volvo 142E          21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
> data<-mtcars
> mean1<-mean(data$wt)
> paste("mean : ",mean1)
[1] "mean :  3.21725"
>
> paste("meadian : ",median(data$wt))
[1] "meadian :  3.325"
>
> paste("Standard Deviation",sd(data$wt))
[1] "Standard Deviation 0.978457442989697"
>
> auto<-mtcars[mtcars$am==0,]
> manual <- mtcars[mtcars$am == 1, ]
>
> t.test(mpg ~ am, data = mtcars)

        Welch Two Sample t-test
```

```
data:  mpg by am
t = -3.7671, df = 18.332, p-value = 0.001374
alternative hypothesis: true difference in means between group 0 and group 1 is not
equal to 0
95 percent confidence interval:
 -11.280194  -3.209684
sample estimates:
mean in group 0 mean in group 1
       17.14737        24.39231

>
> hist(data$hp,xlab="Horse power",main="Histogram of
mtcars",col='blue',border='black',breaks=10)
>
> print("By the given Hypothesis Testin we can say that The average mpg is different
between automatic and manual transmission cars.")
[1] "By the given Hypothesis Testin we can say that The average mpg is different
between automatic and manual transmission cars."
```


Histogram of mtcars
Frequency
0
2
4
6
8
10
50
100
150
200
250
300
350
Horse power

In [3]:
```
import pandas as pd
data=pd.read_csv('C:\\Users\\SHAHNAWAZ\\Downloads\\Employee_data.csv')
```

In [4]:
```
print(data.head())
```

```
   employee_id   firstname    lastname  Department  Salary Joining_date
0          100  Marguerite     Llovera         CSE   76028   25-09-1938
1          101      Phylis        Chem         CSE   40321   10-08-1908
2          102       Renie   Cherianne  Mechanical   45003   11-09-1925
3          103         Ada       Orpah         CSE   38703   09-08-1936
4          104     Caritta  Anastatius         CSE   58067   16-11-1956
```

In [5]:
```
data.isnull().sum()
```

Out[5]:
```
employee_id      0
firstname        0
lastname         0
Department       0
Salary           0
Joining_date     0
dtype: int64
```

In [6]:
```
data.duplicated()
```

Out[6]:
```
0     False
1     False
2     False
3     False
4     False
      ...
95    False
96    False
97    False
98    False
99    False
Length: 100, dtype: bool
```

In [8]:
```
data.drop_duplicates()
```

Out[8]:

	employee_id	firstname	lastname	Department	Salary	Joining_date
0	100	Marguerite	Llovera	CSE	76028	25-09-1938
1	101	Phylis	Chem	CSE	40321	10-08-1908
2	102	Renie	Cherianne	Mechanical	45003	11-09-1925
3	103	Ada	Orpah	CSE	38703	09-08-1936
4	104	Caritta	Anastatius	CSE	58067	16-11-1956
...	...	...	...	...	...	...
95	195	Camile	Evvie	Civil	76506	18-09-2007
96	196	Regina	Aprile	Data Science	87515	03-05-1960
97	197	Blinni	Afton	CSE	46777	23-06-1921
98	198	Kamilah	Stoller	Data Science	18612	19-10-1983
99	199	Daphne	Killigrew	Mechanical	75960	18-01-1901

100 rows × 6 columns

In [10]:
```
data['Joining_date'] = pd.to_datetime(data['Joining_date'],format='%d/%m/%Y')
```

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

```
   employee_id   firstname    lastname  Department  Salary Joining_date
0          100  Marguerite     Llovera         CSE   76028   1938-09-25
1          101      Phylis        Chem         CSE   40321   1908-08-10
2          102       Renie   Cherianne  Mechanical   45003   1925-09-11
3          103         Ada       Orpah         CSE   38703   1936-08-09
4          104     Caritta  Anastatius         CSE   58067   1956-11-16
```

In [15]:
```
average_salary = data.groupby('Department')['Salary'].mean()
print(average_salary)
```

```
Department
CSE              61054.260870
Civil            63196.636364
Data Science     49111.866667
IT               62603.454545
Mechanical       46069.944444
Name: Salary, dtype: float64
```

In [16]:
```
highest_earner = data[data['Salary'] == data['Salary'].max()]
print(highest_earner)
```

```
    employee_id   firstname lastname Department  Salary Joining_date
40          140  Gabriellia     Even      Civil   99642   1927-10-18
```

In [17]:
```
current_year = pd.Timestamp.today().year
data['Years_Worked'] = current_year - data['Joining_date'].dt.year
```

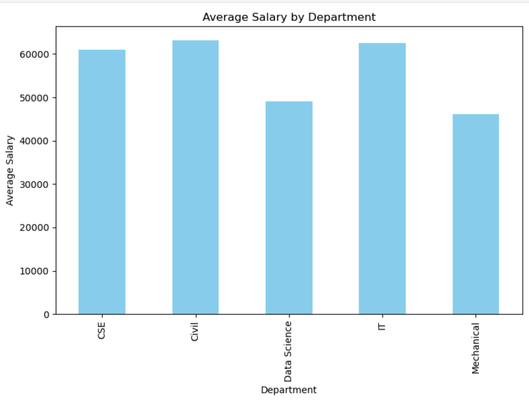
In [21]:
```
print(data)
```

```
    employee_id   firstname    lastname     Department  Salary Joining_date  \
0           100  Marguerite     Llovera            CSE   76028   1938-09-25
1           101      Phylis        Chem            CSE   40321   1908-08-10
2           102       Renie   Cherianne     Mechanical   45003   1925-09-11
3           103         Ada       Orpah            CSE   38703   1936-08-09
4           104     Caritta  Anastatius            CSE   58067   1956-11-16
..          ...         ...         ...            ...     ...          ...
95          195      Camile       Evvie          Civil   76506   2007-09-18
96          196      Regina      Aprile   Data Science   87515   1960-05-03
97          197      Blinni       Afton            CSE   46777   1921-06-23
98          198     Kamilah     Stoller   Data Science   18612   1983-10-19
99          199      Daphne   Killigrew     Mechanical   75960   1901-01-18

    Years_Worked
0             86
1            116
2             99
3             88
4             68
..           ...
95            17
96            64
97           103
98            41
99           123

[100 rows x 7 columns]
```

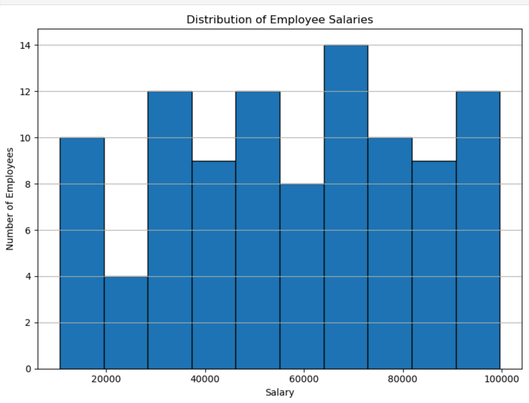
In [22]:
```
average_salary1 = data.groupby('Years_Worked')['Salary'].mean()
print(average_salary1)
```

```
Years_Worked
17     76506.000000
18     58783.500000
20     49291.000000
22     75658.000000
23     50964.666667
           ...
119    97228.000000
121    46859.000000
122    95432.000000
123    75960.000000
124    52423.666667
Name: Salary, Length: 69, dtype: float64
```

In [23]:
```
import matplotlib.pyplot as plt
plt.figure(figsize=(8, 6))
average_salary.plot(kind='bar', color='skyblue')
plt.xlabel('Department')
plt.ylabel('Average Salary')
plt.title('Average Salary by Department')
plt.tight_layout()
plt.show()
```



In [25]:
```
plt.figure(figsize=(8, 6))
plt.hist(data['Salary'], bins=10, edgecolor='black')  # Adjust the number of bins as needed
plt.xlabel('Salary')
plt.ylabel('Number of Employees')
plt.title('Distribution of Employee Salaries')
plt.grid(axis='y')  # Add grid lines for better readability
plt.tight_layout()
plt.show()
```



In [ ]: