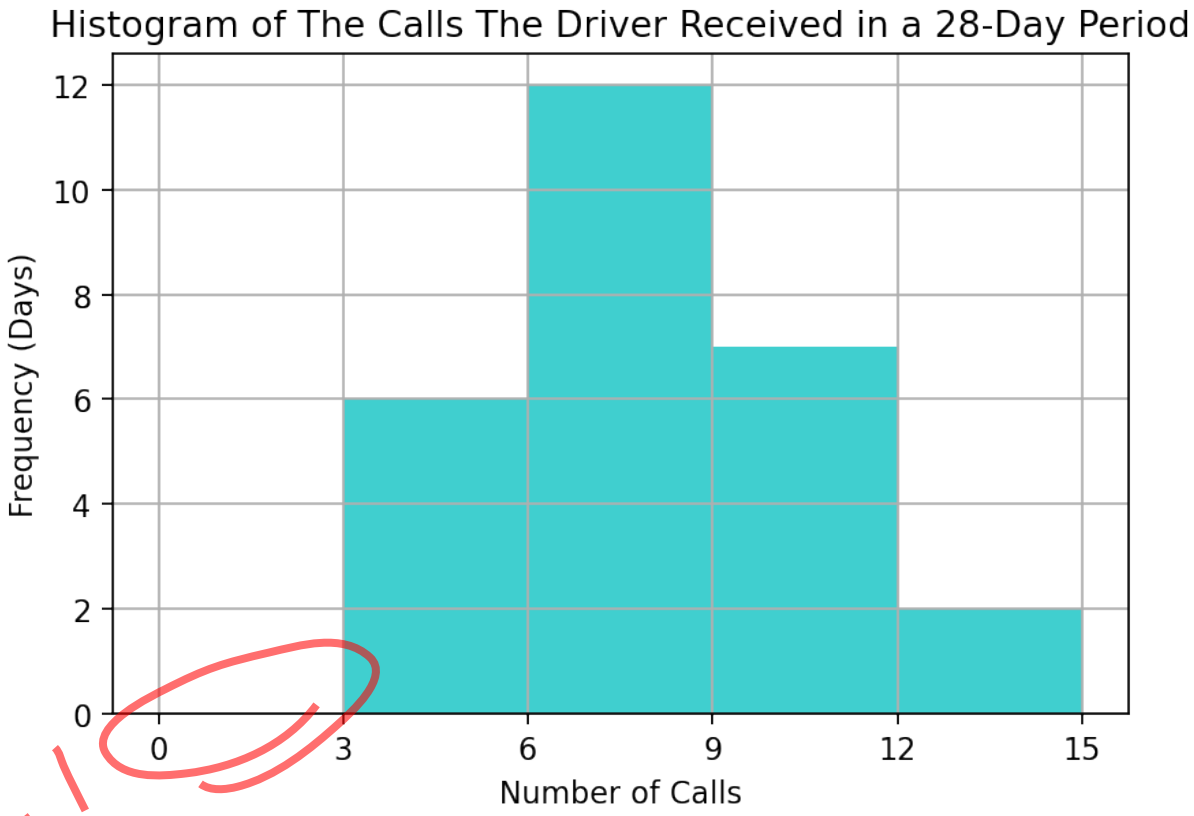
Statistics Homework 02

B08705034 施芊羽 資管二

Chapter 3

3.5

In [2]:
```
#載入所需函式庫
from matplotlib import pyplot as plt
%matplotlib inline
# 設定圖形大小; DPI越大圖越大
plt.rcParams["figure.dpi"] = 160
import seaborn as sns
import pandas as pd
import numpy as np
import scipy.stats as stats
import statsmodels.api as sm
import statsmodels.stats.api as sms
import statsmodels.formula.api as smf
#讀取資料集
df_c03_05 = pd.read_excel('Xr03-05.xlsx')

bins_list = [0, 3, 6, 9, 12, 15]
fig, ax = plt.subplots()
counts, bins, patches = plt.hist(df_c03_05["Calls"], bins=bins_list, den
sity=False, facecolor='c', alpha=0.75)
mu = np.mean(df_c03_05["Calls"])
sigma = np.std(df_c03_05["Calls"])
plt.xlabel('Number of Calls')
plt.ylabel('Frequency (Days)')
plt.title('Histogram of The Calls The Driver Received in a 28-Day Period
')
# plt.text(60, .025, r'$\mu = $', mu,'$\sigma = $', sigma)
plt.grid(True)
plt.xticks(bins_list)

plt.show()
```

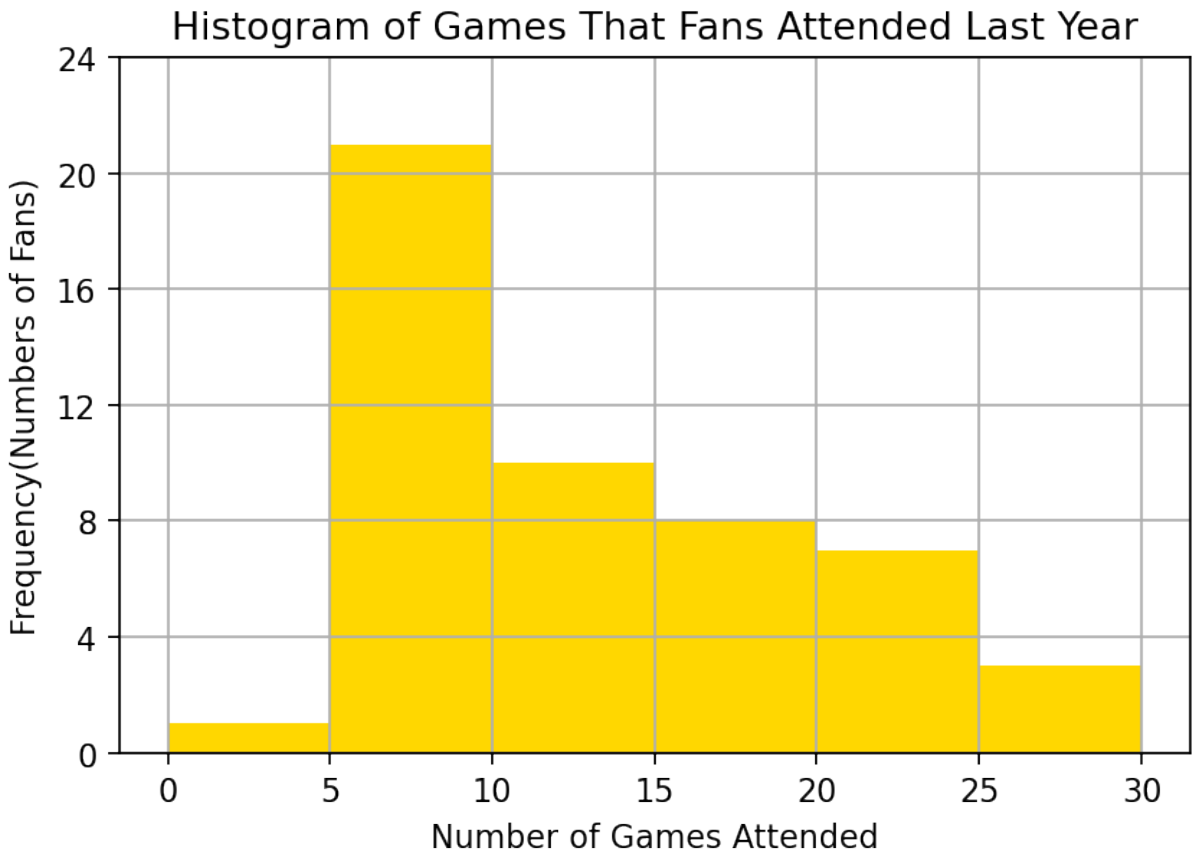


Result:

In this case, we create a histogram and divide the data into five classes. From the graph, we can know the uber driver mostly get 6 − 9 calls per day. Also, we can see that the driver gets more than or equal 3 calls every day.

3.11

In [3]:
```
df_c03_11 = pd.read_excel('Xr03-11.xlsx')

bins_list = [0, 5, 10, 15, 20, 25, 30]
fig, ax = plt.subplots()
counts, bins, patches = plt.hist(df_c03_11["Games"], bins=bins_list, den
sity=False, facecolor='gold', alpha = 1)
mu = np.mean(df_c03_11["Games"])
sigma = np.std(df_c03_11["Games"])
plt.xlabel('Number of Games Attended')
plt.ylabel('Frequency(Numbers of Fans)')
plt.title('Histogram of Games That Fans Attended Last Year')
# plt.text(60, .025, r'$\mu = $', mu,'$\sigma = $', sigma)
plt.grid(True)
plt.yticks(np.arange(0, 25, 4), np.arange(0, 25, 4))
plt.xticks(bins_list)

plt.show()
```

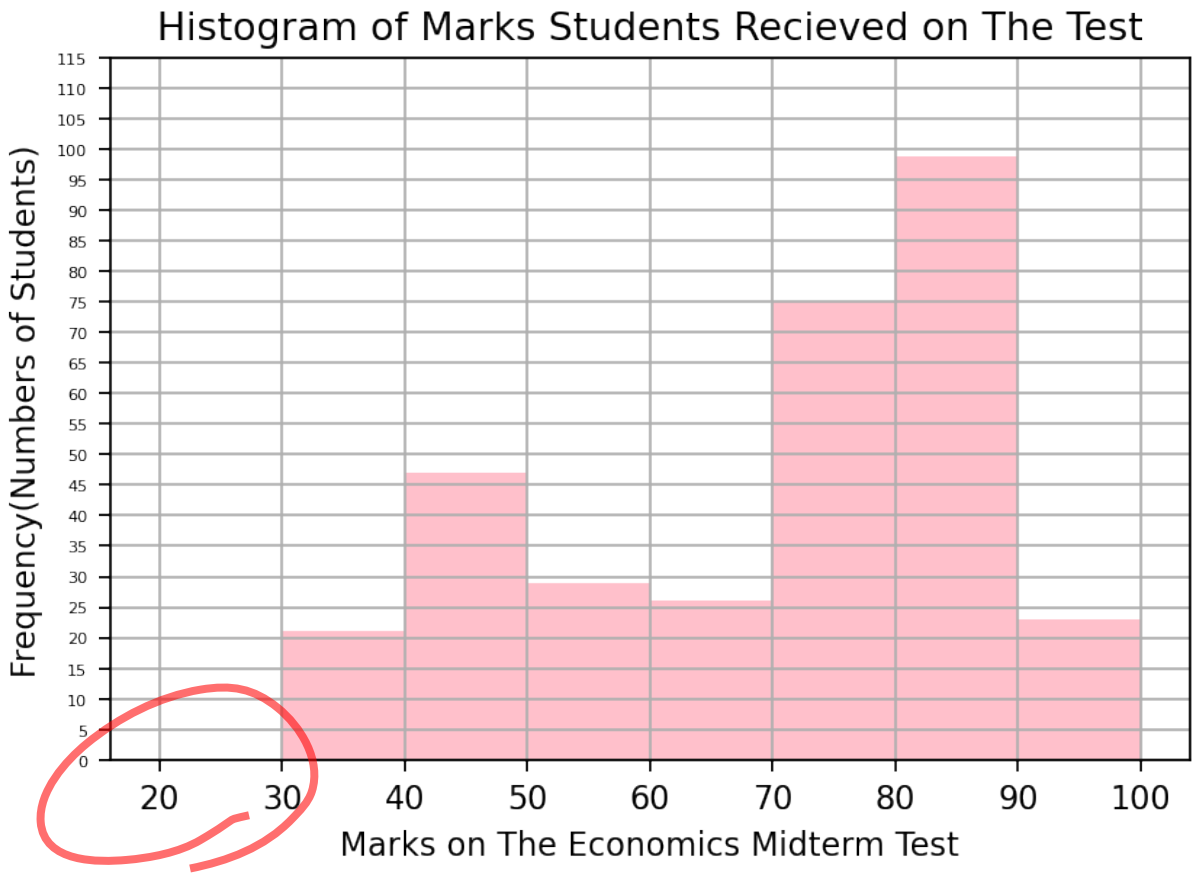


Result:

After making the histogram with 6 classes, we can see that the histogram is **unimodal and negatively skewed**.

3.17

In [4]:
```
df_c03_17 = pd.read_excel('Xr03-17.xlsx')

bins_list = [20, 30, 40, 50, 60, 70, 80, 90, 100]
fig, ax = plt.subplots()
counts, bins, patches = plt.hist(df_c03_17["Marks"], bins=bins_list, den
sity=False, facecolor='pink', alpha = 1)
mu = np.mean(df_c03_17["Marks"])
sigma = np.std(df_c03_17["Marks"])
plt.xlabel('Marks on The Economics Midterm Test')
plt.ylabel('Frequency(Numbers of Students)')
plt.title('Histogram of Marks Students Recieved on The Test')
# plt.text(60, .025, r'$\mu = $', mu,'$\sigma = $', sigma)
plt.grid(True)
plt.yticks(np.arange(0, 120, 5), np.arange(0, 120, 5), fontsize=5)
plt.xticks(bins_list)

plt.show()
```



Result:

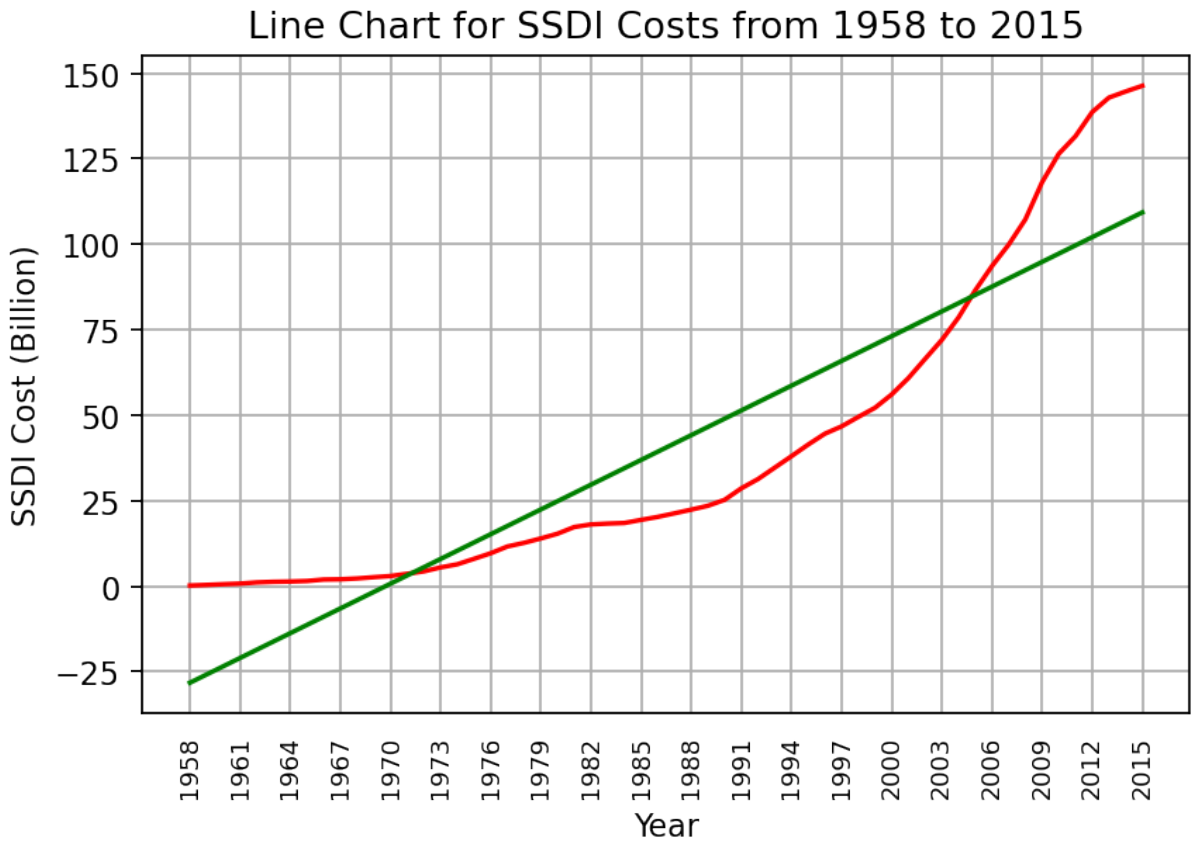
Since the graph is **bimodal**, we can see that most students get a mark around 70 − 90, and in the rest of students, most of them got a score from 40 to 50. Hence, we can say that the test might have a good discrimination rate because students can get a good score in it but hard to get higher (over 90), and it also can testify a student's preparation on the test since there are still a part of students failed on it.

3.39

a.

In [3]:
```
df_c03_39 = pd.read_excel('Xr03-39.xlsx')


plt.plot(df_c03_39["Year"], df_c03_39["Disability Ins (DI) -fed"] , colo
r='red')

plt.tick_params(
    axis='x',          # changes apply to the x-axis
    which='both',      # both major and minor ticks are affected
    bottom=False,      # ticks along the bottom edge are off
    top=False,         # ticks along the top edge are off
    )
plt.xticks(np.arange(1958, 2020, 3), np.arange(1958, 2020, 3), rotation
= 90, fontsize=8)
# plt.yticks(np.arange(0, 150, 10), np.arange(0, 150, 10), fontsize=8)
plt.subplots_adjust(bottom=0.15)
```

```
plt.xlabel('Year')
plt.ylabel('SSDI Cost (Billion)')
plt.grid(True)
m, b = np.polyfit(df_c03_39["Year"], df_c03_39["Disability Ins (DI) -fed
"], 1) # line of best fit
plt.plot(df_c03_39["Year"], m*df_c03_39["Year"] + b, color = 'g')
plt.title('Line Chart for SSDI Costs from 1958 to 2015')
plt.show()
```



Result:

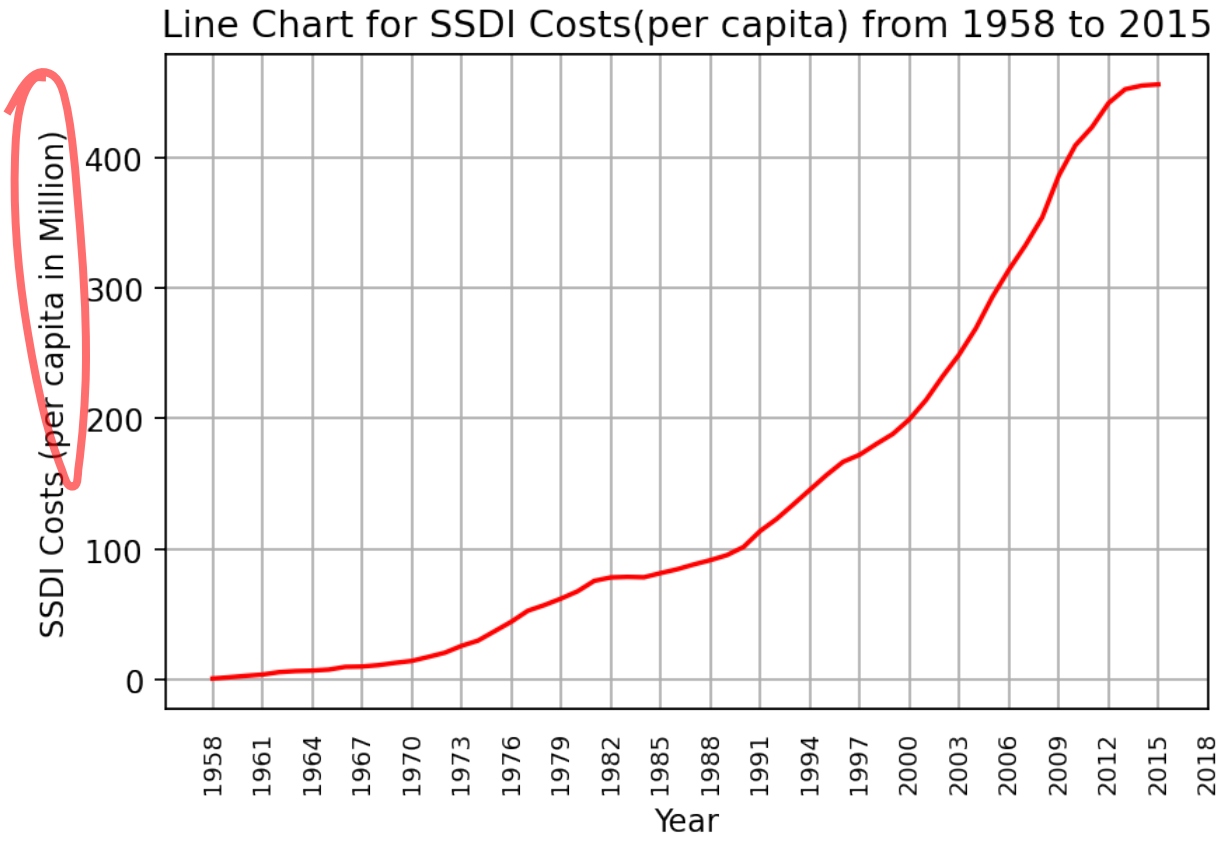
From the line chart, we can see that the SSDI cost has been increasing since 1958 in the original data. That means the government took an eye on Social Security Disability Insurance more. **The data doesn't show the unit of it, so I compare other exercise and guess that's in billion.**

great!

b.

In [4]:
```
population = pd.read_excel("U.S. Population 1935-2015.xlsx")
y = np.arange(2015 - 1958 + 1)

for i in range(2015 - 1958 + 1):
    y[i] = population["U.S. Population (millions)"][i + 1958 - 1935]

plt.plot(df_c03_39["Year"], df_c03_39["Disability Ins (DI) -fed"] / y *
1000 , color='red')

plt.tick_params(
    axis='x',          # changes apply to the x-axis
    which='both',      # both major and minor ticks are affected
    bottom=False,      # ticks along the bottom edge are off
    top=False,         # ticks along the top edge are off
```

```
    )
k = {0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50}
plt.xticks(np.arange(1958, 2020, 3), np.arange(1958, 2020, 3), rotation
= 90, fontsize=8)
plt.subplots_adjust(bottom=0.15)
plt.xlabel('Year')
plt.ylabel('SSDI Costs (per capita in Million)')
plt.grid(True)
plt.title('Line Chart for SSDI Costs(per capita) from 1958 to 2015')
plt.show()
```



Result:

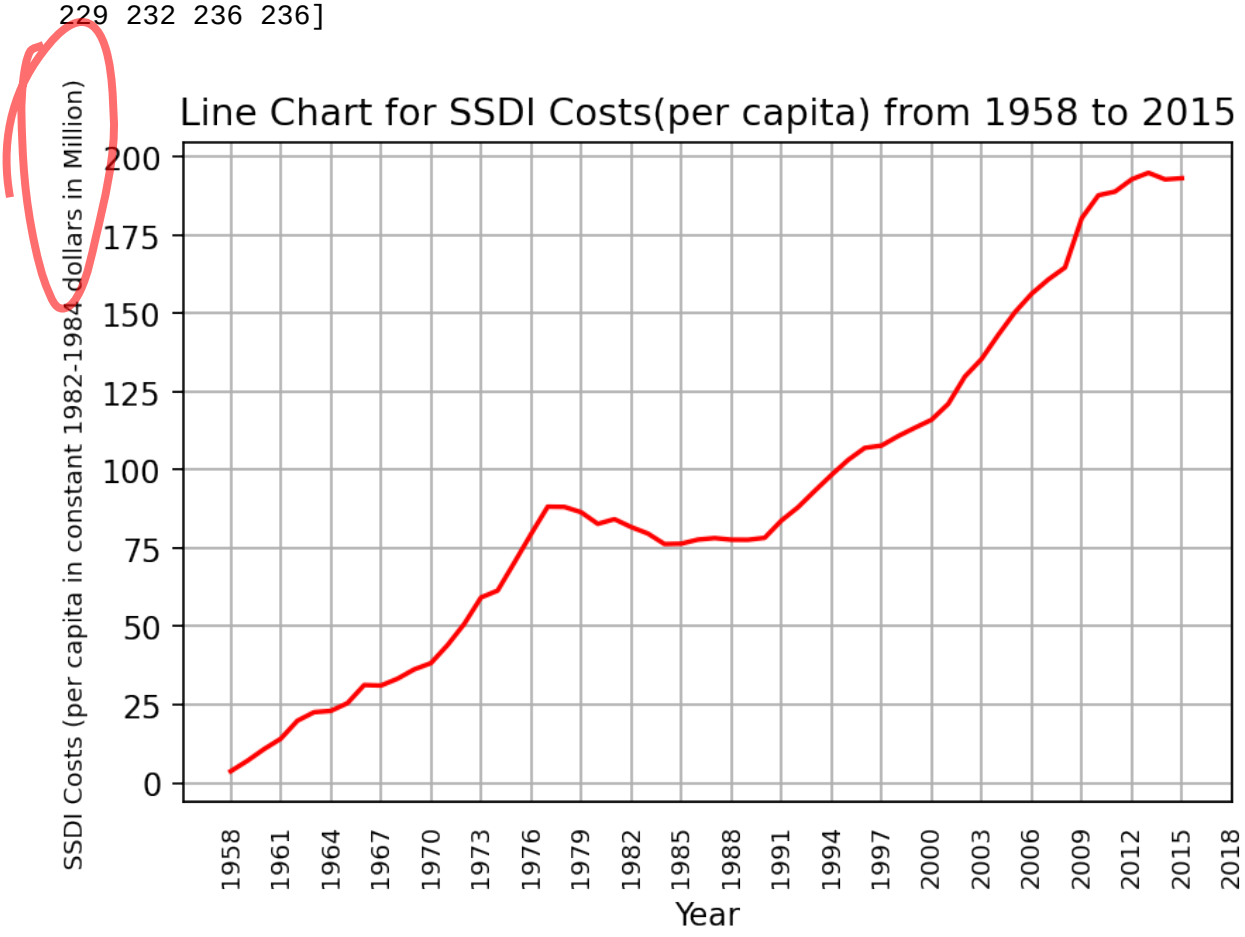
At first, I change the unit from billion to million. In the graph, we can see that the SSDI cost per capita had been growing since 1958.

C.

In [5]:
```
cpi = pd.read_excel("U.S. CPI Annual.xlsx")
c = np.arange(2015 - 1958 + 1)

for i in range(2015 - 1958 + 1):
    c[i] = cpi["All Urban Consumers - (CPI-U): U.S. city average: All it
ems: 1982-84=100"][i + 1958 - 1913]

plt.plot(df_c03_39["Year"], df_c03_39["Disability Ins (DI) -fed"] / y *
100000 / c , color='red') # Divide by population and set in constant 1982-1984 dollars

plt.tick_params(
    axis='x',          # changes apply to the x-axis
    which='both',      # both major and minor ticks are affected
```

```
    bottom=False,        # ticks along the bottom edge are off
    top=False,           # ticks along the top edge are off
    )
k = {0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50}
plt.xticks(np.arange(1958, 2020, 3), np.arange(1958, 2020, 3), rotation
= 90, fontsize=8)
plt.subplots_adjust(bottom=0.15)
plt.xlabel('Year')
plt.ylabel('SSDI Costs (per capita in constant 1982-1984 dollars in Mill
ion)', fontsize=8)
plt.grid(True)
plt.title('Line Chart for SSDI Costs(per capita) from 1958 to 2015')
plt.show()
```

```
[ 28  29  29  29  30  30  31  31  32  33  34  36  38  40  41  44  49  53
  56  60  65  72  82  90  96  99 103 107 109 113 118 123 130 136 140 144
 148 152 156 160 163 166 172 177 179 184 188 195 201 207 215 214 218 224
 229 232 236 236]
```



Result:

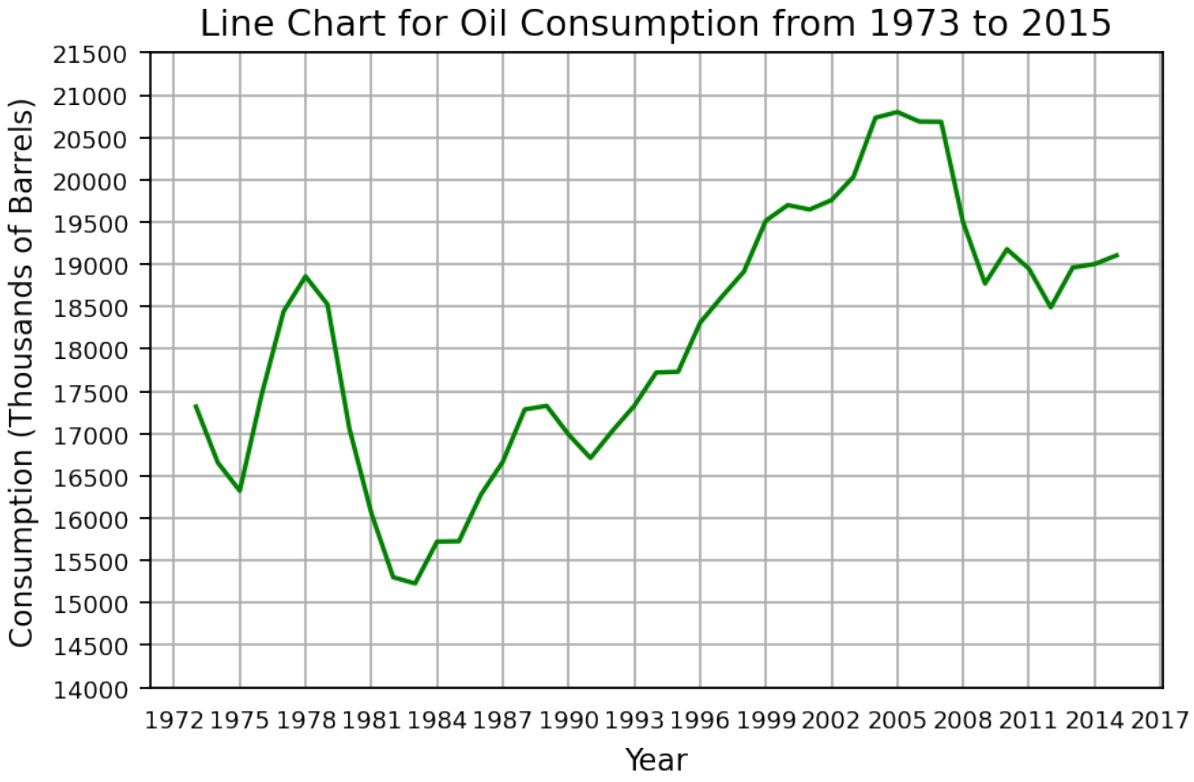
From the graph we can know that the SSDI cost per capita in constant 1982-1984 dollars kept increasing from 1958 except for a small decreasing from 1977 to 1984.

To conclude, even in different presentation of SSDI cost, the graphs all show a growth trend.

3.47

In [33]:
```
df_c03_47 = pd.read_excel('Xr03-47.xlsx')

plt.plot(df_c03_47["Year"], df_c03_47["Consumption"], color='g')
```

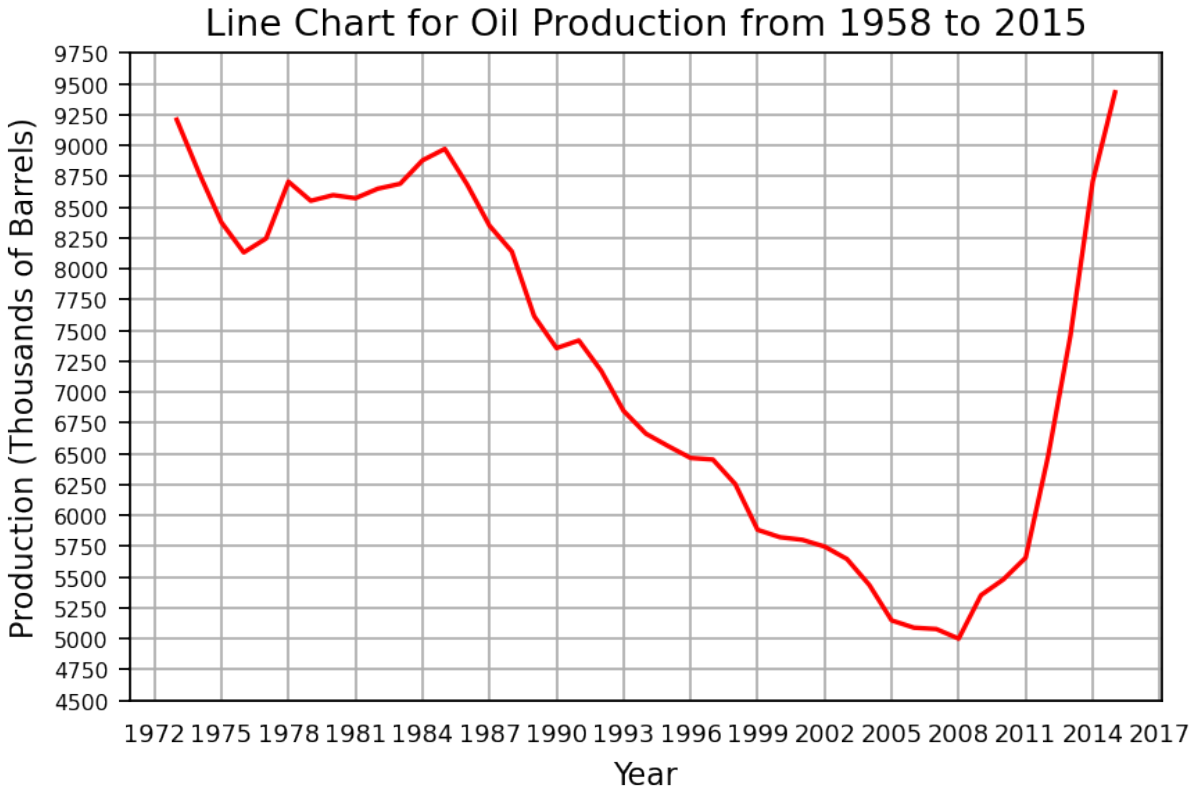
```
plt.tick_params(
    axis='x',          # changes apply to the x-axis
    which='both',      # both major and minor ticks are affected
    bottom=False,      # ticks along the bottom edge are off
    top=False,         # ticks along the top edge are off
    )

plt.xticks(np.arange(1972, 2020, 3), np.arange(1972, 2020, 3), fontsize=
8)
plt.yticks(np.arange(14000, 22000, 500), np.arange(14000, 22000, 500), f
ontsize=8)
plt.subplots_adjust(bottom=0.15)
plt.xlabel('Year')
plt.ylabel('Consumption (Thousands of Barrels)')
plt.grid(True)
plt.title('Line Chart for Oil Consumption from 1973 to 2015')
plt.show()
```



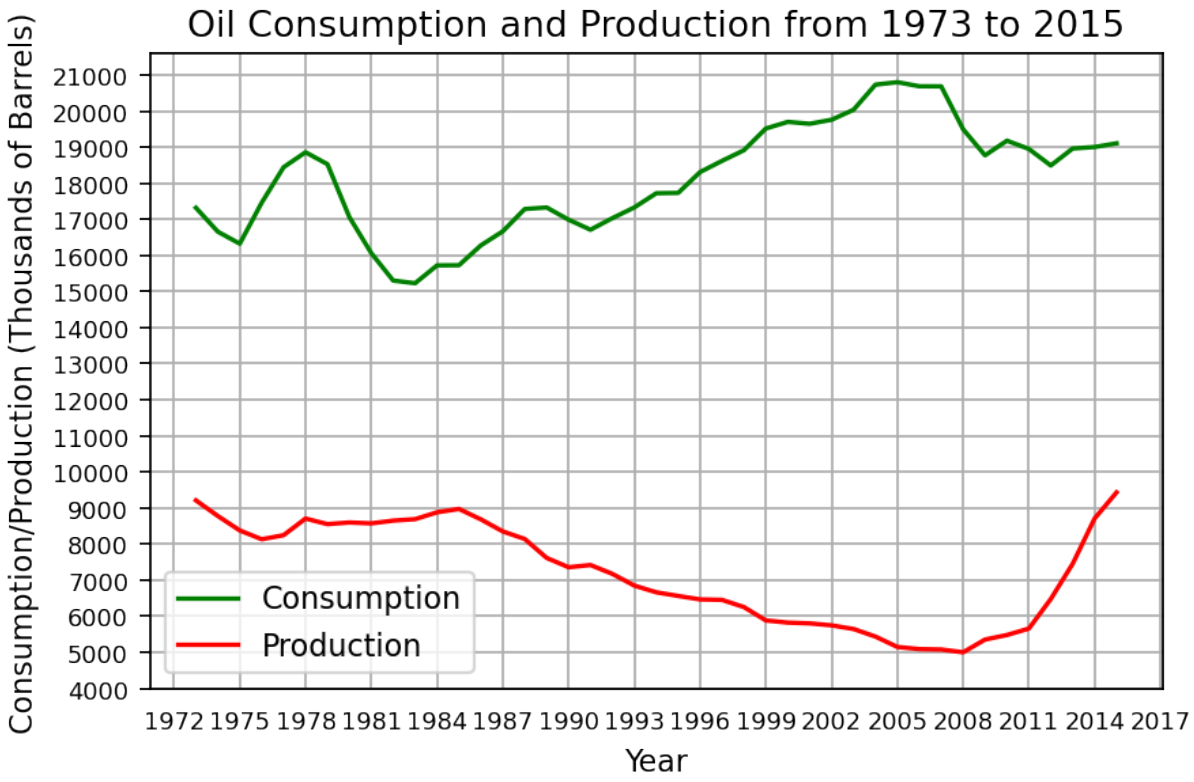
In [72]:

```
plt.plot(df_c03_47["Year"], df_c03_47["Production"], color='r')

plt.tick_params(
    axis='x',          # changes apply to the x-axis
    which='both',      # both major and minor ticks are affected
    bottom=False,      # ticks along the bottom edge are off
    top=False,         # ticks along the top edge are off
    )
plt.xticks(np.arange(1972, 2020, 3), np.arange(1972, 2020, 3), fontsize=
8)
plt.yticks(np.arange(4500, 10000, 250), np.arange(4500, 10000, 250), fon
tsize=7)
plt.subplots_adjust(bottom=0.15)
plt.xlabel('Year')
plt.ylabel('Production (Thousands of Barrels)')
plt.grid(True)
plt.title('Line Chart for Oil Production from 1958 to 2015')
```

```
plt.show()
```



In [34]:
```
plt.plot(df_c03_47["Year"], df_c03_47["Consumption"], color='g', label =
 "Consumption")

plt.tick_params(
    axis='x',          # changes apply to the x-axis
    which='both',      # both major and minor ticks are affected
    bottom=False,      # ticks along the bottom edge are off
    top=False,         # ticks along the top edge are off
    )

plt.plot(df_c03_47["Year"], df_c03_47["Production"], color='r', label =
"Production")
plt.tick_params(
    axis='x',          # changes apply to the x-axis
    which='both',      # both major and minor ticks are affected
    bottom=False,      # ticks along the bottom edge are off
    top=False,         # ticks along the top edge are off
    )
plt.xticks(np.arange(1972, 2020, 3), np.arange(1972, 2020, 3), fontsize=
8)
plt.yticks(np.arange(4000, 22000, 1000), np.arange(4000, 22000, 1000), f
ontsize=8)
plt.subplots_adjust(bottom=0.15)
plt.xlabel('Year')
plt.ylabel('Consumption/Production (Thousands of Barrels)')
plt.grid(True)
plt.legend()
plt.title('Oil Consumption and Production from 1973 to 2015')
plt.show()
```



Result:

To see more clearly toward the data, I've made three graphs which display oil consumption only, oil production only, and both on the three charts.

We can see that the oil consumption grew a bit from 1975 to 1978 and suffered a small fall since 1978. Ever since 1983, the oil consumption had been mostly increasing year by year, in my own opinion, I think it might be caused by the more uses of cars and other transportation. Around 2007, it started to fall and became more stable. I think that's because of the global awareness of global warming and the improvement in technology which made the vehicles more energy-saving.

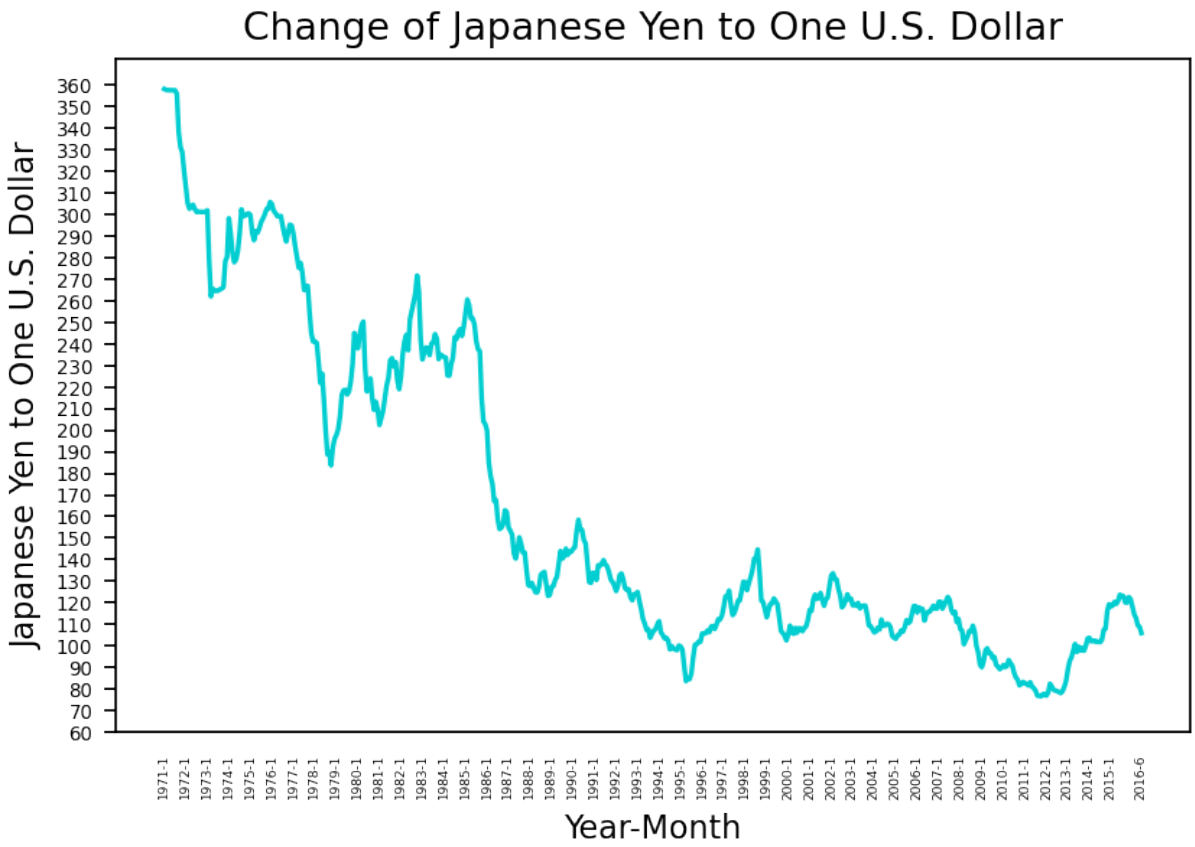
As of oil production, we can see that it had been decreasing since 1985, I think one of the reason might be that the oil resource had been mostly discovered and used, hence, the oil production decreased slightly year after year. In 2008, the growth may also be caused from the technology invention which gave us more way to produce more oil.

We can find that oil consumption is always more than oil production, which makes us aware of the oil resource, because as the way we consumed oil, the oil might be went out one day.

3.53

In [35]:
```
df_c03_53 = pd.read_excel('Xr03-53.xlsx')

n = df_c03_53["Year"].size
year = np.array(df_c03_53["Year"])
month = np.array(df_c03_53["Month"])
y_m = [" " for k in range(n)]
for i in range(45):
    y_m[i*12] = str(year[i*12]) + "-" + str(month[i*12])
y_m[n - 1] = str(year[n -1]) + "-" + str(month[n - 1])
```

```
df_c03_53["time"] = (df_c03_53["Year"] - 1971) * 12 + df_c03_53["Month"]

plt.plot(df_c03_53["time"], df_c03_53["Japanese Yen to One U.S. Dollar"]
, color='darkturquoise')
plt.xticks(df_c03_53["time"], y_m, rotation=90, fontsize=4)
plt.tick_params(
    axis='x',          # changes apply to the x-axis
    which='both',      # both major and minor ticks are affected
    bottom=False,      # ticks along the bottom edge are off
    top=False,         # ticks along the top edge are off
    )

plt.yticks(np.arange(60, 370, 10), np.arange(60, 370, 10), fontsize=6)
plt.subplots_adjust(bottom=0.15)
plt.xlabel('Year-Month')
plt.ylabel('Japanese Yen to One U.S. Dollar')

plt.title('Change of Japanese Yen to One U.S. Dollar')
plt.show()
```

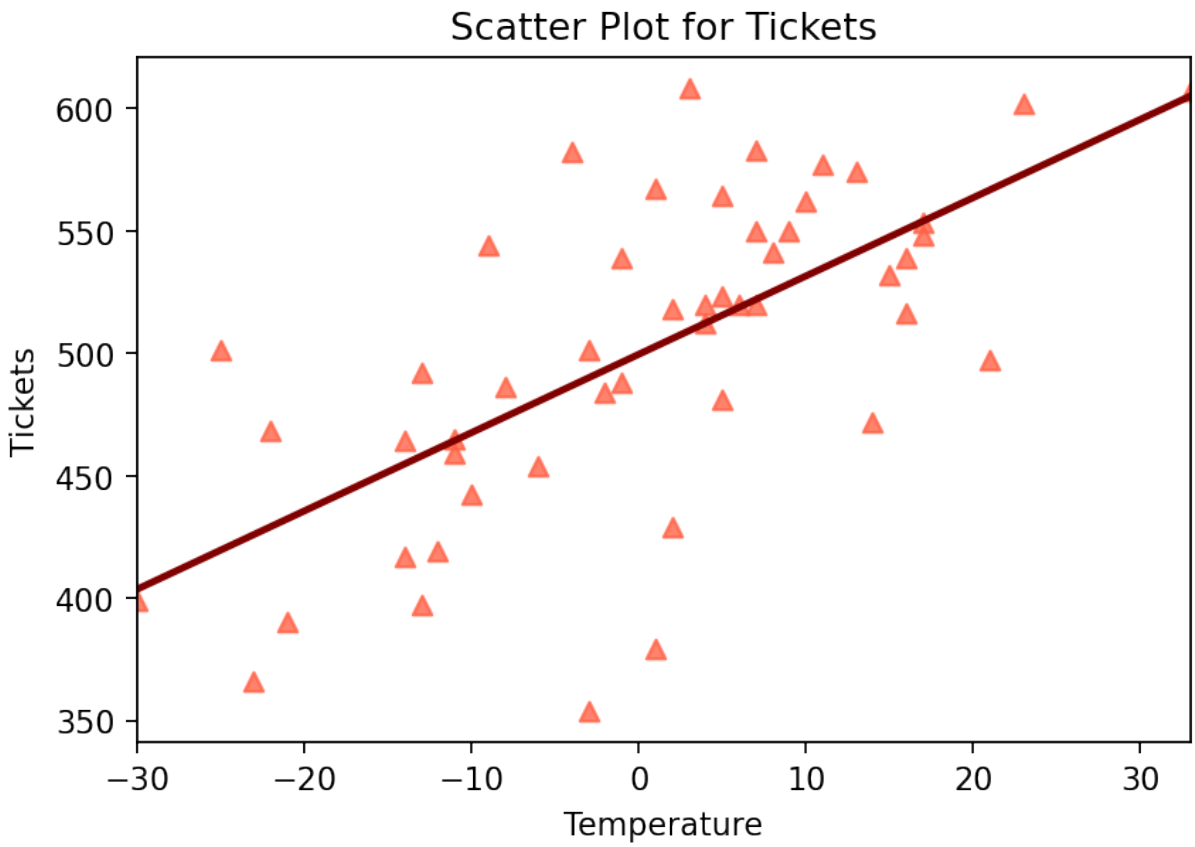


Result:

In the graph, I first intepret the change rate of Japanese Yen to one US dollar was highest (around 360) in 1971. The trend of the change kept changing every year, I think it's because of the change may be caused by the world situation. World keeps changing every day, so the trend of the changes may not always be the same, it will go up and down due to several changes in the world. In 1985, it decreased a lot I think it might be caused of the Plaza Accord. Ever since that year, Japanese Yen to one US dollar had never gone higher than 170.

3.63

In [36]:
```
df_c03_63 = pd.read_excel('Xr03-63.xlsx')

plt.title('Scatter Plot for Tickets')
plt.xlabel('Temperature')
plt.ylabel('Tickets')
# plt.show()

_ = sns.regplot(x='Temperature', y= 'Tickets', data = df_c03_63, scatter
_kws={"color": "tomato"}, marker="^", line_kws={"color": "maroon"}, ci =
 None)
plt.title('Scatter Plot for Tickets')
plt.show()
```



Result:

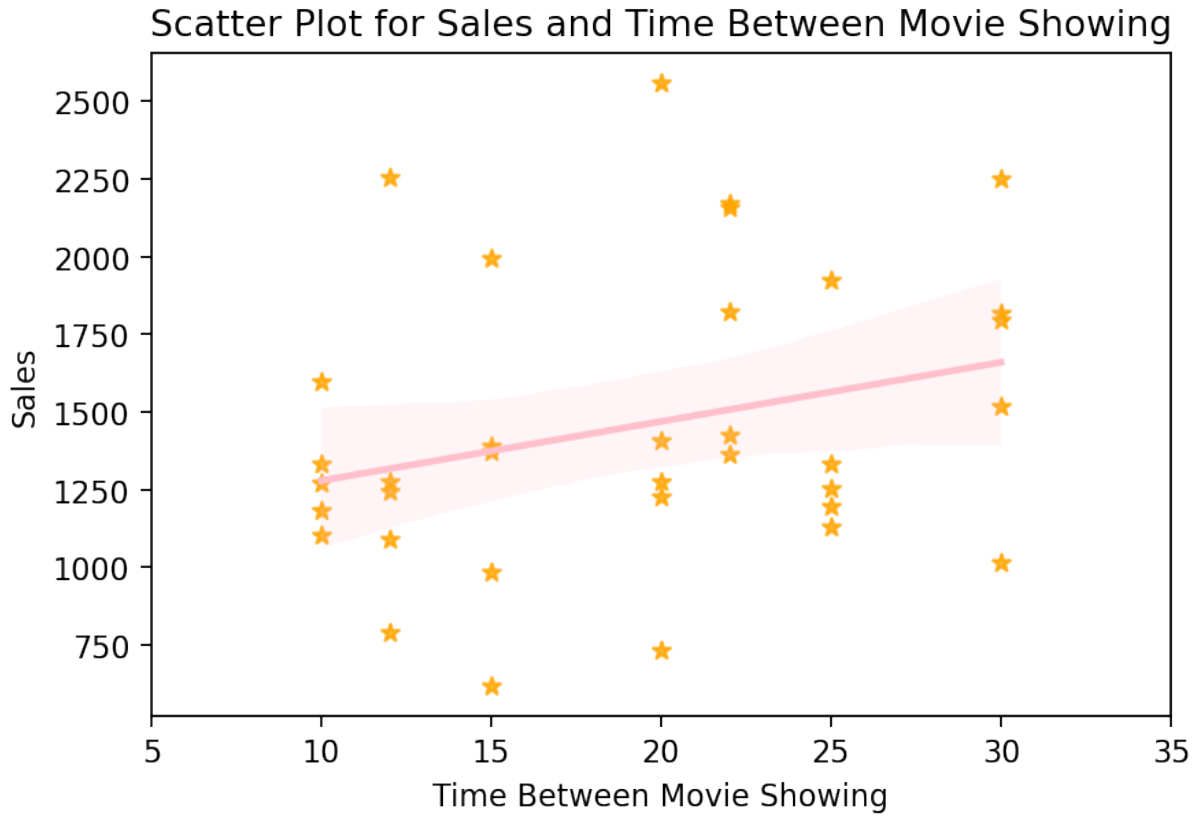
From the line of best fit, we can see that mostly in a colder weather less tickets might be sold. I think that if the weather is too cold may make skiers less desired of going skiing. Also, most randomly selected days had a temperature around -10 to 10 degree. In conclusion, I'd like to say that the temperature is mostly aroud -10 to 10 degree in the winter months, and more people go skiing on a day that the temperature is not that low.

3.67

a.

In [38]:
```
df_c03_67 = pd.read_excel('Xr03-67.xlsx')

_ = sns.regplot(x='Time', y= 'Sales', data = df_c03_67, scatter_kws={"co
lor": "orange"}, marker="*", line_kws={"color": "pink"})
```

```
plt.title('Scatter Plot for Sales and Time Between Movie Showing')
plt.xticks(np.arange(5, 40, 5), np.arange(5, 40, 5))
plt.xlabel('Time Between Movie Showing')
plt.show()
```



b.

Even though the line of trend shows that there's a positive relationship, from the scatter points, it's still hard to determine whether the relationship between the two variables is positive or not.

3.83

a.

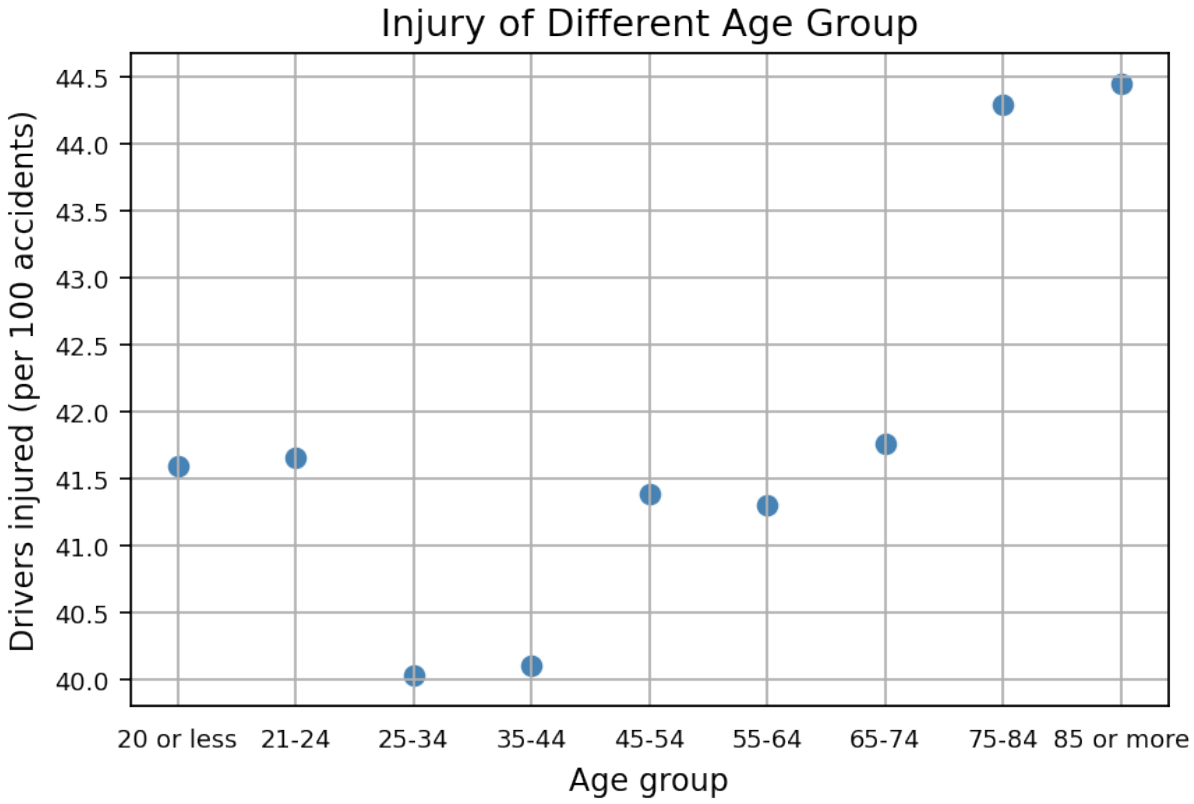
In [42]:
```
df_c03_83 = pd.read_excel('Xr03-83.xlsx')

df_c03_83['Drivers injury rate(per 100 accidents)'] = df_c03_83['Drivers
 injured']/df_c03_83['Number of accidents'] * 100.00
df_c03_83['Drivers death rate(per accident)'] =  df_c03_83['Drivers kill
ed']/df_c03_83['Number of accidents']
display(df_c03_83)
```

	Age group	Number of accidents	Drivers injured	Drivers killed	Drivers injury rate(per 100 accidents)	Drivers death rate(per accident)
0	20 or less	52313	21762	217	41.599602	0.004148
1	21-24	38449	16016	185	41.655180	0.004812
2	25-34	78703	31503	324	40.027699	0.004117

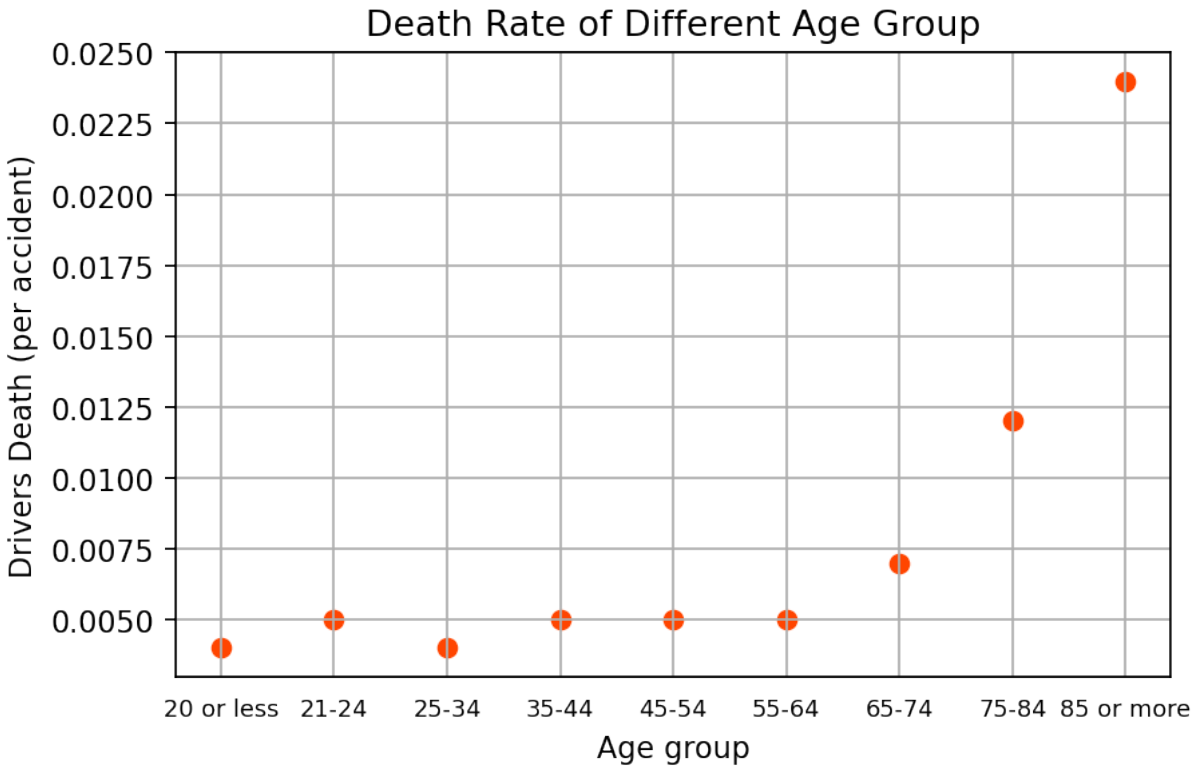
3	35-44	76152	30542	389	40.106629	0.005108
4	45-54	54699	22638	260	41.386497	0.004753
5	55-64	31985	13210	167	41.300610	0.005221
6	65-74	18896	7892	133	41.765453	0.007039
7	75-84	11526	5106	138	44.299844	0.011973
8	85 or more	2751	1223	65	44.456561	0.023628

b.

In [47]:
```
sizes = df_c03_83['Drivers injury rate(per 100 accidents)'] # Collect the data of each country
labels =  df_c03_83['Age group']

plt.scatter(df_c03_83["Age group"], np.around(df_c03_83["Drivers injury rate(per 100 accidents)"], 3), color='steelblue', label = "Drivers injury rate")
plt.tick_params(
    axis='x',           # changes apply to the x-axis
    which='both',       # both major and minor ticks are affected
    bottom=False,       # ticks along the bottom edge are off
    top=False,          # ticks along the top edge are off
    )
plt.xticks(labels, fontsize=8, rotation = 0)
plt.yticks(np.arange(40, 45, 0.5), np.arange(40, 45, 0.5), fontsize=8)
plt.subplots_adjust(bottom=0.15)
plt.xlabel('Age group')
plt.ylabel('Drivers injured (per 100 accidents)')
plt.grid(True)
plt.title('Injury of Different Age Group')
plt.show()
```



In [46]:
```
sizes = df_c03_83['Drivers death rate(per accident)'] # Collect the data of each c
ountry
labels =  df_c03_83['Age group']

plt.scatter(df_c03_83["Age group"], np.around(df_c03_83["Drivers death r
ate(per accident)"], 3), color='orangered', label = "Drivers death rate"
)
plt.tick_params(
    axis='x',          # changes apply to the x-axis
    which='both',      # both major and minor ticks are affected
    bottom=False,      # ticks along the bottom edge are off
    top=False,         # ticks along the top edge are off
    )

plt.xticks(labels, fontsize=8, rotation = 0)
plt.subplots_adjust(bottom=0.15)
plt.xlabel('Age group')
plt.ylabel('Drivers Death (per accident)')
plt.grid(True)
plt.title('Death Rate of Different Age Group')
plt.show()
```



c.

From the graphs above, we can know that the injury rate isn't affected by age a lot. The lowest rate occured in the age group from 25 to 44. However, for those age higher than 65, the injury rate becomes a lot more higher, I think that's because aging might affect on the immunication and health. In my personal opinion, those younger than 25 years old have a higher injury rate may be caused by they are not so mature both in mental and physical health.

As for death rate, we can see that it's more associated with age. We can directly see from the graph that those in higher age group have a higher death rate.

The death rate is far more lower than the injury rate for sure; in conclusion, we can say that death rate has a closer relationship to age than injury rate, and for both rates, the elders (over 65 years old) suffered a higher rates than the younger groups.