Load data

In [1]:
```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_r
emount=True).
```

In [2]:
```
path = "/content/drive/My Drive"
path
```

Out[2]:
```
'/content/drive/My Drive'
```

In [3]:
```
import pandas as pd
import seaborn as sns
import statistics
from subprocess import check_output
from scipy import stats
import matplotlib.pyplot as plt
```

In [4]:
```
data = pd.read_csv(path +"/cust_seg.csv")
```

In [5]:
```
data.head()
```

Out[5]:

	custid	sex	AqChannel	region	Marital_status	segment	pre_usage	Post_usage_1month	Latest_mon_usage	post_usage_2ndmonth
0	70	0	4	1	1	1	57	52	49.2	57.2
1	121	1	4	2	1	3	68	59	63.6	64.9
2	86	0	4	3	1	1	44	33	64.8	36.3
3	141	0	4	3	1	3	63	44	56.4	48.4
4	172	0	4	2	1	2	47	52	68.4	57.2

In [6]:
```
data.columns
```

Out[6]:
```
Index(['custid', 'sex', 'AqChannel', 'region', 'Marital_status', 'segment',
       'pre_usage', 'Post_usage_1month', 'Latest_mon_usage',
       'post_usage_2ndmonth'],
      dtype='object')
```

In [7]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 10 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   custid               200 non-null    int64  
 1   sex                  200 non-null    int64  
 2   AqChannel            200 non-null    int64  
 3   region               200 non-null    int64  
 4   Marital_status       200 non-null    int64  
 5   segment              200 non-null    int64  
 6   pre_usage            200 non-null    int64  
 7   Post_usage_1month    200 non-null    int64  
 8   Latest_mon_usage     200 non-null    float64
 9   post_usage_2ndmonth  200 non-null    float64
dtypes: float64(2), int64(8)
memory usage: 15.8 KB
```

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

Out[8]:

	custid	sex	AqChannel	region	Marital_status	segment	pre_usage	Post_usage_1month	Latest_mon_usage	post_usa
count	200.000000	200.00000	200.000000	200.000000	200.000000	200.000000	200.000000	200.000000	200.000000	
mean	100.500000	0.54500	3.430000	2.055000	1.160000	2.025000	52.230000	52.775000	63.174000	
std	57.879185	0.49922	1.039472	0.724291	0.367526	0.690477	10.252937	9.478586	11.242137	
min	1.000000	0.00000	1.000000	1.000000	1.000000	1.000000	28.000000	31.000000	39.600000	
25%	50.750000	0.00000	3.000000	2.000000	1.000000	2.000000	44.000000	45.750000	54.000000	
50%	100.500000	1.00000	4.000000	2.000000	1.000000	2.000000	50.000000	54.000000	62.400000	
75%	150.250000	1.00000	4.000000	3.000000	1.000000	2.250000	60.000000	60.000000	70.800000	
max	200.000000	1.00000	4.000000	3.000000	2.000000	3.000000	76.000000	67.000000	90.000000	

In [9]:
```
data.Latest_mon_usage.mean()
```

Out[9]:
```
63.17400000000001
```

In [10]:
```
data.Latest_mon_usage.std()
```

Out[10]:
```
11.242137352892753
```

Hypothesis Testing

One Sample T-test

Card usage has been improved from last year usage which was 50 One sample t-test

H0 (Null hypothesis) Sample_avg = 50

Ha (ALT hypothesis) Sample_avg > 50

If this is P < 0.05 and Sample_avg > 50 satisfied we can reject the NULL hypothesis. else we fail to reject the NULL hypothesis.

In [11]:
```
stats.ttest_1samp(a = data.Latest_mon_usage, popmean = 50) # pop mean is the hypothetical value
```

Out[11]:
```
Ttest_1sampResult(statistic=16.57233752433133, pvalue=2.4963719280931583e-39)
```

In [12]:
```
data.Latest_mon_usage.mean()
```

Out[12]:
```
63.17400000000001
```

Since the p value is very small we reject the NULL hypothesis.

Hence we reject the NULL hypotheiss and accept the ALT hypothesis

#Two Sample T-Test(Paired)

The last campaign was successful in terms of credit card.

We can compare pre_usage and post_usage of the credit card - Paired sample t-test(dependent sample t-test)

H0 (NULL hypothesis) Pre_avg = post_avg

Ha (ALT hypothesis) pre_avg < post_avg

if p < 0.05 and pre_avg < post_avg, you will reject null

In [13]:
```
#Two Sample T-Test(Paired)
print(data.pre_usage.mean())
print(data.Post_usage_1month.mean())
print(data.post_usage_2ndmonth.mean())
```

```
52.23
52.775
58.05250000000003
```

In [14]:
```
stats.ttest_rel(a = data.pre_usage, b = data.Post_usage_1month)
```

Out[14]:
```
Ttest_relResult(statistic=-0.8673065458794775, pvalue=0.3868186820914985)
```

Since the pvalue is not less than 0.05 we fail to reject the Null Hypothesis. Hence the campaign was successful

Two sample T-Test(Independent)

Is there any differce in credit card spend usage between males and females? since there are only two categories we can do independent sample t test here.

Comparing two sample averages (both are independent samples)

H0: males_avg = females_avg

Ha: males_avg <> females_avg

if p<0.05, then we reject NULL (there is a relationship between sex and spend)

else, you fail to reject the NULL (There is no relationship between sex and spend)

In [15]:
```
Males_spend = data.Post_usage_1month[data.sex == 0]
Females_spend = data.Post_usage_1month[data.sex == 1]
```

In [16]:
```
print(Males_spend.head())
print(Females_spend.head())
```

```
0    52
2    33
3    44
4    52
5    52
Name: Post_usage_1month, dtype: int64
1     59
92    62
93    44
94    44
95    62
Name: Post_usage_1month, dtype: int64
```

In [17]:
```
print(Males_spend.mean())
print(Females_spend.mean())
```

```
50.120879120879124
54.99082568807339
```

In [18]:
```
print(Males_spend.std())
print(Females_spend.std())
```

```
10.305160697259263
8.13371516959346
```

In [19]:
```
stats.ttest_ind(a = Males_spend, b = Females_spend, equal_var = False)
```

Out[19]:
```
Ttest_indResult(statistic=-3.6564080478875276, pvalue=0.00034088493594266187)
```

Since pvalue is less than 0.05 we reject the NULL hypothesis. Hence, we conclude there is a difference between male spend and female spend

chi-squred Test

Is there any relationship between region and segment? Chi Square test

H0 : There is no relationship

Ha : There is relationship

if p < 0.05 , then we reject NULL,

else, we fail to reject NULL Hypothesis.

In [20]:
```
t = pd.crosstab(data.segment, data.region, margins = True)
t
```

Out[20]:

segment	region 1	2	3	All
1	16	20	9	45
2	19	44	42	105
3	12	31	7	50
All	47	95	58	200

In [21]:
```
chi2, p, dof, expected = stats.chi2_contingency(t)
print("P-value :" ,p)
print("chi-squared statistic :",chi2)
print("Degree of freedom :" , dof)
print("Expected value :" , expected)
```

```
P-value : 0.055282939487992365
chi-squared statistic : 16.60444164948934
Degree of freedom : 9
Expected value : [[ 10.575  21.375  13.05   45.   ]
 [ 24.675  49.875  30.45  105.   ]
 [ 11.75   23.75   14.5    50.   ]
 [ 47.     95.     58.    200.   ]]
```

Based on the p value we can say there is a relationship between region and segment

#Z-Test

In [26]:
```
df = pd.read_csv(path +"/blood_pressure.csv")
```

In [29]:
```
df.head()
```

Out[29]:

	patient	sex	agegrp	bp_before	bp_after
0	1	Male	30-45	143	153
1	2	Male	30-45	163	170
2	3	Male	30-45	153	168
3	4	Male	30-45	153	142
4	5	Male	30-45	146	141

one-sample Z test

z-test for blood pressure with some mean like 156

In [27]:
```
ztest ,pval = stests.ztest(df['bp_before'], x2=None, value=156)
print(float(pval))
if pval<0.05:
    print("reject null hypothesis")
else:
    print("accept null hypothesis")
```

```
0.6651614730255063
accept null hypothesis
```

Two-sample Z test

H0 : mean of two group is 0

H1 : mean of two group is not 0

Example : we are checking in blood data after blood and before blood data.

In [31]:
```
ztest ,pval = stests.ztest(df['bp_before'], x2=df['bp_after'], value=0,alternative='two-sided')
print(float(pval))
if pval<0.05:
    print("reject null hypothesis")
else:
    print("accept null hypothesis")
```

```
0.002162306611369422
reject null hypothesis
```