In [1]:

```
import pandas as pd
import numpy as np
from sklearn import preprocessing
```

In [2]:

```
df = pd.read_csv('Placement_Data_Full_Class.csv')
```

In [3]:

```
df
```

Out[3]:

| | sl_no | gender | ssc_p | ssc_b | hsc_p | hsc_b | hsc_s | degree_p | degree_t | worl |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | M | 67.00 | Others | 91.00 | Others | Commerce | 58.00 | Sci&Tech | |
| **1** | 2 | M | 79.33 | Central | 78.33 | Others | Science | 77.48 | Sci&Tech | |
| **2** | 3 | M | 65.00 | Central | 68.00 | Central | Arts | 64.00 | Comm&Mgmt | |
| **3** | 4 | M | 56.00 | Central | 52.00 | Central | Science | 52.00 | Sci&Tech | |
| **4** | 5 | M | 85.80 | Central | 73.60 | Central | Commerce | 73.30 | Comm&Mgmt | |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **210** | 211 | M | 80.60 | Others | 82.00 | Others | Commerce | 77.60 | Comm&Mgmt | |
| **211** | 212 | M | 58.00 | Others | 60.00 | Others | Science | 72.00 | Sci&Tech | |
| **212** | 213 | M | 67.00 | Others | 67.00 | Others | Commerce | 73.00 | Comm&Mgmt | |
| **213** | 214 | F | 74.00 | Others | 66.00 | Others | Commerce | 58.00 | Comm&Mgmt | |
| **214** | 215 | M | 62.00 | Central | 58.00 | Others | Science | 53.00 | Comm&Mgmt | |

215 rows × 15 columns

In [4]:

```
df.shape
```

Out[4]:

```
(215, 15)
```

In [5]:

```
print("Rows = ",df.shape[0])
print("Columns = ",df.shape[1])
```

```
Rows =  215
Columns =  15
```

In [6]:

```
df.dtypes
```

Out[6]:

```
sl_no              int64
gender            object
ssc_p            float64
ssc_b             object
hsc_p            float64
hsc_b             object
hsc_s             object
degree_p         float64
degree_t          object
workex            object
etest_p          float64
specialisation    object
mba_p            float64
status            object
salary           float64
dtype: object
```

In [7]:

```
df.head()
```

Out[7]:

| | sl_no | gender | ssc_p | ssc_b | hsc_p | hsc_b | hsc_s | degree_p | degree_t | worke |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | M | 67.00 | Others | 91.00 | Others | Commerce | 58.00 | Sci&Tech | N |
| **1** | 2 | M | 79.33 | Central | 78.33 | Others | Science | 77.48 | Sci&Tech | Ye |
| **2** | 3 | M | 65.00 | Central | 68.00 | Central | Arts | 64.00 | Comm&Mgmt | N |
| **3** | 4 | M | 56.00 | Central | 52.00 | Central | Science | 52.00 | Sci&Tech | N |
| **4** | 5 | M | 85.80 | Central | 73.60 | Central | Commerce | 73.30 | Comm&Mgmt | N |

In [8]:

```
df.tail()
```

Out[8]:

| | sl_no | gender | ssc_p | ssc_b | hsc_p | hsc_b | hsc_s | degree_p | degree_t | work |
|---|---|---|---|---|---|---|---|---|---|---|
| **210** | 211 | M | 80.6 | Others | 82.0 | Others | Commerce | 77.6 | Comm&Mgmt | |
| **211** | 212 | M | 58.0 | Others | 60.0 | Others | Science | 72.0 | Sci&Tech | |
| **212** | 213 | M | 67.0 | Others | 67.0 | Others | Commerce | 73.0 | Comm&Mgmt | Y |
| **213** | 214 | F | 74.0 | Others | 66.0 | Others | Commerce | 58.0 | Comm&Mgmt | |
| **214** | 215 | M | 62.0 | Central | 58.0 | Others | Science | 53.0 | Comm&Mgmt | |

In [9]:

```
df.describe()
```

Out[9]:

| | sl_no | ssc_p | hsc_p | degree_p | etest_p | mba_p | sala |
|---|---|---|---|---|---|---|---|
| **count** | 215.000000 | 215.000000 | 215.000000 | 215.000000 | 215.000000 | 215.000000 | 148.0000 |
| **mean** | 108.000000 | 67.303395 | 66.333163 | 66.370186 | 72.100558 | 62.278186 | 288655.4054 |
| **std** | 62.209324 | 10.827205 | 10.897509 | 7.358743 | 13.275956 | 5.833385 | 93457.4524 |
| **min** | 1.000000 | 40.890000 | 37.000000 | 50.000000 | 50.000000 | 51.210000 | 200000.0000 |
| **25%** | 54.500000 | 60.600000 | 60.900000 | 61.000000 | 60.000000 | 57.945000 | 240000.0000 |
| **50%** | 108.000000 | 67.000000 | 65.000000 | 66.000000 | 71.000000 | 62.000000 | 265000.0000 |
| **75%** | 161.500000 | 75.700000 | 73.000000 | 72.000000 | 83.500000 | 66.255000 | 300000.0000 |
| **max** | 215.000000 | 89.400000 | 97.700000 | 91.000000 | 98.000000 | 77.890000 | 940000.0000 |

In [10]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 215 entries, 0 to 214
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   sl_no           215 non-null    int64  
 1   gender          215 non-null    object 
 2   ssc_p           215 non-null    float64
 3   ssc_b           215 non-null    object 
 4   hsc_p           215 non-null    float64
 5   hsc_b           215 non-null    object 
 6   hsc_s           215 non-null    object 
 7   degree_p        215 non-null    float64
 8   degree_t        215 non-null    object 
 9   workex          215 non-null    object 
 10  etest_p         215 non-null    float64
 11  specialisation  215 non-null    object 
 12  mba_p           215 non-null    float64
 13  status          215 non-null    object 
 14  salary          148 non-null    float64
dtypes: float64(6), int64(1), object(8)
memory usage: 25.3+ KB
```

In [11]:

```
df.mean(numeric_only = True)
```

Out[11]:

```
sl_no          108.000000
ssc_p           67.303395
hsc_p           66.333163
degree_p        66.370186
etest_p         72.100558
mba_p           62.278186
salary      288655.405405
dtype: float64
```

In [12]:

```
df.max(numeric_only = True)
```

Out[12]:

```
sl_no          215.00
ssc_p           89.40
hsc_p           97.70
degree_p        91.00
etest_p         98.00
mba_p           77.89
salary      940000.00
dtype: float64
```

In [13]:

```
df.min(numeric_only = True)
```

Out[13]:

```
sl_no            1.00
ssc_p           40.89
hsc_p           37.00
degree_p        50.00
etest_p         50.00
mba_p           51.21
salary      200000.00
dtype: float64
```

In [14]:

```
df.median()
```

```
C:\Users\shrey\AppData\Local\Temp\ipykernel_22352\530051474.py:1: FutureW
arning: Dropping of nuisance columns in DataFrame reductions (with 'numer
ic_only=None') is deprecated; in a future version this will raise TypeErr
or.  Select only valid columns before calling the reduction.
  df.median()
```

Out[14]:

```
sl_no          108.0
ssc_p           67.0
hsc_p           65.0
degree_p        66.0
etest_p         71.0
mba_p           62.0
salary      265000.0
dtype: float64
```

In [15]:

```
df.mode(numeric_only = True)
```

Out[15]:

| | sl_no | ssc_p | hsc_p | degree_p | etest_p | mba_p | salary |
|---|---|---|---|---|---|---|---|
| **0** | 1 | 62.0 | 63.0 | 65.0 | 60.0 | 56.7 | 300000.0 |
| **1** | 2 | NaN | NaN | NaN | NaN | NaN | NaN |
| **2** | 3 | NaN | NaN | NaN | NaN | NaN | NaN |
| **3** | 4 | NaN | NaN | NaN | NaN | NaN | NaN |
| **4** | 5 | NaN | NaN | NaN | NaN | NaN | NaN |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **210** | 211 | NaN | NaN | NaN | NaN | NaN | NaN |
| **211** | 212 | NaN | NaN | NaN | NaN | NaN | NaN |
| **212** | 213 | NaN | NaN | NaN | NaN | NaN | NaN |
| **213** | 214 | NaN | NaN | NaN | NaN | NaN | NaN |
| **214** | 215 | NaN | NaN | NaN | NaN | NaN | NaN |

215 rows × 7 columns

In [16]:

```
df.columns
```

Out[16]:

```
Index(['sl_no', 'gender', 'ssc_p', 'ssc_b', 'hsc_p', 'hsc_b', 'hsc_s',
       'degree_p', 'degree_t', 'workex', 'etest_p', 'specialisation', 'mb
a_p',
       'status', 'salary'],
      dtype='object')
```

In [17]:

```
df.isnull()
```

Out[17]:

| | sl_no | gender | ssc_p | ssc_b | hsc_p | hsc_b | hsc_s | degree_p | degree_t | workex | etest |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | False | False | False | False | False | False | False | False | False | False | Fa |
| **1** | False | False | False | False | False | False | False | False | False | False | Fa |
| **2** | False | False | False | False | False | False | False | False | False | False | Fa |
| **3** | False | False | False | False | False | False | False | False | False | False | Fa |
| **4** | False | False | False | False | False | False | False | False | False | False | Fa |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **210** | False | False | False | False | False | False | False | False | False | False | Fa |
| **211** | False | False | False | False | False | False | False | False | False | False | Fa |
| **212** | False | False | False | False | False | False | False | False | False | False | Fa |
| **213** | False | False | False | False | False | False | False | False | False | False | Fa |
| **214** | False | False | False | False | False | False | False | False | False | False | Fa |

215 rows × 15 columns

In [18]:

```
df.isna().sum()
```

Out[18]:

```
sl_no             0
gender            0
ssc_p             0
ssc_b             0
hsc_p             0
hsc_b             0
hsc_s             0
degree_p          0
degree_t          0
workex            0
etest_p           0
specialisation    0
mba_p             0
status            0
salary           67
dtype: int64
```

In [19]:

```
df.isna().sum().sum()
```

Out[19]:

```
67
```

In [20]:

```
print(df['ssc_p'].min())
print(df['ssc_p'].max())
print(df['ssc_p'].std())
print(df['ssc_p'].quantile(0.25))
print(df['ssc_p'].quantile(0.5))
print(df['ssc_p'].count())
```

```
40.89
89.4
10.827205398231452
60.599999999999994
67.0
215
```

In [21]:

```
df.dtypes
```

Out[21]:

```
sl_no              int64
gender            object
ssc_p            float64
ssc_b             object
hsc_p            float64
hsc_b             object
hsc_s             object
degree_p         float64
degree_t          object
workex            object
etest_p          float64
specialisation    object
mba_p            float64
status            object
salary           float64
dtype: object
```

In [22]:

```
df['gender'].astype(str)
```

Out[22]:

```
0      M
1      M
2      M
3      M
4      M
      ..
210    M
211    M
212    M
213    F
214    M
Name: gender, Length: 215, dtype: object
```

In [23]:

```
df.dtypes
```

Out[23]:

```
sl_no              int64
gender            object
ssc_p            float64
ssc_b             object
hsc_p            float64
hsc_b             object
hsc_s             object
degree_p         float64
degree_t          object
workex            object
etest_p          float64
specialisation    object
mba_p            float64
status            object
salary           float64
dtype: object
```

In [24]:

```
df1 = df
```

In [25]:

```
df1
```

Out[25]:

| | sl_no | gender | ssc_p | ssc_b | hsc_p | hsc_b | hsc_s | degree_p | degree_t | worl |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | M | 67.00 | Others | 91.00 | Others | Commerce | 58.00 | Sci&Tech | |
| **1** | 2 | M | 79.33 | Central | 78.33 | Others | Science | 77.48 | Sci&Tech | |
| **2** | 3 | M | 65.00 | Central | 68.00 | Central | Arts | 64.00 | Comm&Mgmt | |
| **3** | 4 | M | 56.00 | Central | 52.00 | Central | Science | 52.00 | Sci&Tech | |
| **4** | 5 | M | 85.80 | Central | 73.60 | Central | Commerce | 73.30 | Comm&Mgmt | |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **210** | 211 | M | 80.60 | Others | 82.00 | Others | Commerce | 77.60 | Comm&Mgmt | |
| **211** | 212 | M | 58.00 | Others | 60.00 | Others | Science | 72.00 | Sci&Tech | |
| **212** | 213 | M | 67.00 | Others | 67.00 | Others | Commerce | 73.00 | Comm&Mgmt | |
| **213** | 214 | F | 74.00 | Others | 66.00 | Others | Commerce | 58.00 | Comm&Mgmt | |
| **214** | 215 | M | 62.00 | Central | 58.00 | Others | Science | 53.00 | Comm&Mgmt | |

215 rows × 15 columns

In [26]:

```
df1['hsc_p'] = df1['hsc_p'].astype(str)
```

In [27]:

```
df1.dtypes
```

Out[27]:

```
sl_no              int64
gender            object
ssc_p            float64
ssc_b             object
hsc_p             object
hsc_b             object
hsc_s             object
degree_p         float64
degree_t          object
workex            object
etest_p          float64
specialisation    object
mba_p            float64
status            object
salary           float64
dtype: object
```

In [28]:

```
df1['hsc_p'] = df1['hsc_p'].astype(float)
```

In [29]:

```
df1.isna().sum()
```

Out[29]:

```
sl_no             0
gender            0
ssc_p             0
ssc_b             0
hsc_p             0
hsc_b             0
hsc_s             0
degree_p          0
degree_t          0
workex            0
etest_p           0
specialisation    0
mba_p             0
status            0
salary           67
dtype: int64
```

In [30]:

```
df1['salary'] = df1['salary'].fillna(df['salary'].mean())
```

In [31]:

```
df1.isna().sum()
```

Out[31]:

```
sl_no             0
gender            0
ssc_p             0
ssc_b             0
hsc_p             0
hsc_b             0
hsc_s             0
degree_p          0
degree_t          0
workex            0
etest_p           0
specialisation    0
mba_p             0
status            0
salary            0
dtype: int64
```

In [34]:

```
label = preprocessing.LabelEncoder()
for column in df1:
    if df1[column].dtypes == "object":
        df1[column] = label.fit_transform(df1[column])
```

In [35]:

```
df1.dtypes
```

Out[35]:

```
sl_no              int64
gender             int32
ssc_p            float64
ssc_b              int32
hsc_p            float64
hsc_b              int32
hsc_s              int32
degree_p         float64
degree_t           int32
workex             int32
etest_p          float64
specialisation     int32
mba_p            float64
status             int32
salary           float64
dtype: object
```

In [47]:

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
df1 = scaler.fit_transform(df1)
df1 = pd.DataFrame(df1,columns =df.columns)
df1
```

Out[47]:

| | sl_no | gender | ssc_p | ssc_b | hsc_p | hsc_b | hsc_s | degree_p | degree_t | worke |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0.000000 | 1.0 | 0.538240 | 1.0 | 0.889621 | 1.0 | 0.5 | 0.195122 | 1.0 | 0 |
| **1** | 0.004673 | 1.0 | 0.792414 | 0.0 | 0.680890 | 1.0 | 1.0 | 0.670244 | 1.0 | 1 |
| **2** | 0.009346 | 1.0 | 0.497011 | 0.0 | 0.510708 | 0.0 | 0.0 | 0.341463 | 0.0 | 0 |
| **3** | 0.014019 | 1.0 | 0.311482 | 0.0 | 0.247117 | 0.0 | 1.0 | 0.048780 | 1.0 | 0 |
| **4** | 0.018692 | 1.0 | 0.925788 | 0.0 | 0.602965 | 0.0 | 0.5 | 0.568293 | 0.0 | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **210** | 0.981308 | 1.0 | 0.818594 | 1.0 | 0.741351 | 1.0 | 0.5 | 0.673171 | 0.0 | 0 |
| **211** | 0.985981 | 1.0 | 0.352711 | 1.0 | 0.378913 | 1.0 | 1.0 | 0.536585 | 1.0 | 0 |
| **212** | 0.990654 | 1.0 | 0.538240 | 1.0 | 0.494234 | 1.0 | 0.5 | 0.560976 | 0.0 | 1 |
| **213** | 0.995327 | 0.0 | 0.682540 | 1.0 | 0.477759 | 1.0 | 0.5 | 0.195122 | 0.0 | 0 |
| **214** | 1.000000 | 1.0 | 0.435168 | 0.0 | 0.345964 | 1.0 | 1.0 | 0.073171 | 0.0 | 0 |

215 rows × 15 columns

In [45]:

```
df1.columns
```

Out[45]:

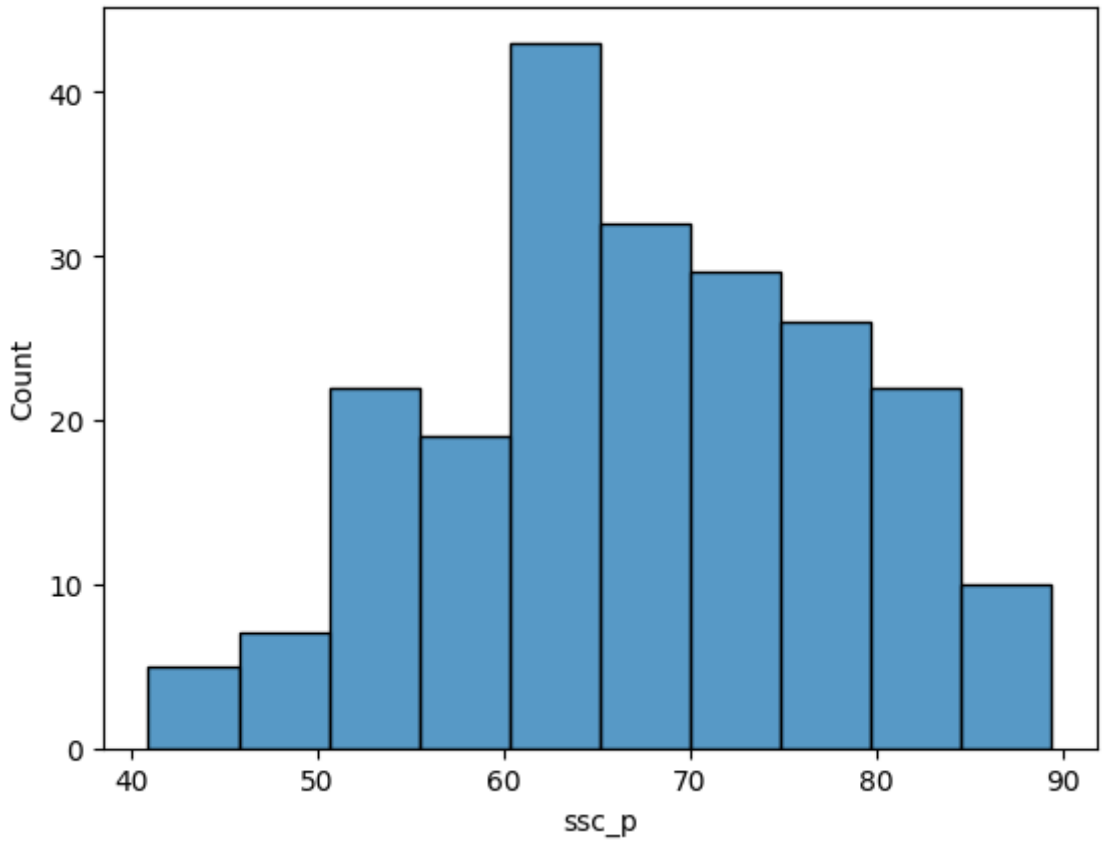
```
Index(['sl_no', 'gender', 'ssc_p', 'ssc_b', 'hsc_p', 'hsc_b', 'hsc_s',
       'degree_p', 'degree_t', 'workex', 'etest_p', 'specialisation', 'mb
a_p',
       'status', 'salary'],
      dtype='object')
```

In [49]:

```
import seaborn as sns
sns.histplot(data = df,x = 'ssc_p')
```

Out[49]:

```
<Axes: xlabel='ssc_p', ylabel='Count'>
```




In [ ]: