In [1]:
```
import pandas as pd
import seaborn as sb

fileName = 'Absenteeism_at_work.xls'
filePath = '/Users/tomisin/Dropbox/My Mac (Tomisins-MacBook-Pro.local)/Documents/MY WORKSPACE/Absenteeism_at_work/'
data = pd.read_excel(filePath + fileName)
```

In [2]:
```
'''Using the Absenteeism dataset, conduct exploratory data analysis and satisfy the following requirements;

1. Describe the behavior of the dataset. Focus on Missing values, Duplicates, Shape, and features behavior

*** From .isna().sum(), there were no missing values at all.

*** There were 34 duplicate rows which were dropped altogether.

*** From .shape, the data shape contains 740 samples with 21 columns.

*** Observing from .describe(), this dataset needs to the normalized
because the majority of the data in the labels
are not normally distributed.









'''
```

Out[2]:
```
'Using the Absenteeism dataset, conduct exploratory data analysis and satisfy the following requirements;\n\n1. Describe
the behavior of the dataset. Focus on Missing values, Duplicates, Shape, and features behavior\n\n'
```

In [122…
```
data.shape
```

Out[122…
```
(740, 21)
```

In [3]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 740 entries, 0 to 739
Data columns (total 21 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   ID                               740 non-null    int64
 1   Reason for absence               740 non-null    int64
 2   Month of absence                 740 non-null    int64
 3   Day of the week                  740 non-null    int64
 4   Seasons                          740 non-null    int64
 5   Transportation expense           740 non-null    int64
 6   Distance from Residence to Work  740 non-null    int64
 7   Service time                     740 non-null    int64
 8   Age                              740 non-null    int64
 9   Work load Average/day            740 non-null    int64
 10  Hit target                       740 non-null    int64
 11  Disciplinary failure             740 non-null    int64
 12  Education                        740 non-null    int64
 13  Son                              740 non-null    int64
 14  Social drinker                   740 non-null    int64
 15  Social smoker                    740 non-null    int64
 16  Pet                              740 non-null    int64
 17  Weight                           740 non-null    int64
 18  Height                           740 non-null    int64
 19  Body mass index                  740 non-null    int64
 20  Absenteeism time in hours        740 non-null    int64
dtypes: int64(21)
memory usage: 121.5 KB
```

In [4]:

```
data
```

Out[4]:

	ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work	Service time	Age	Work load Average/day	...	Disciplinary failure	Education	Son	Social drinker	s
0	11	26	7	3	1	289	36	13	33	239554	...	0	1	2	1	
1	36	0	7	3	1	118	13	18	50	239554	...	1	1	1	1	
2	3	23	7	4	1	179	51	18	38	239554	...	0	1	0	1	
3	7	7	7	5	1	279	5	14	39	239554	...	0	1	2	1	
4	11	23	7	5	1	289	36	13	33	239554	...	0	1	2	1	

	ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work	Service time	Age	Work load Average/day	...	Disciplinary failure	Education	Son	Social drinker	s
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
735	11	14	7	3	1	289	36	13	33	264604	...	0	1	2	1	
736	1	11	7	3	1	235	11	14	37	264604	...	0	3	1	0	
737	4	0	0	3	1	118	14	13	40	271219	...	0	1	1	1	
738	8	0	0	4	2	231	35	14	39	271219	...	0	1	2	1	
739	35	0	0	6	3	179	45	14	53	271219	...	0	1	1	0	

740 rows × 21 columns

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

Out[5]:
```
ID                               0
Reason for absence               0
Month of absence                 0
Day of the week                  0
Seasons                          0
Transportation expense           0
Distance from Residence to Work  0
Service time                     0
Age                              0
Work load Average/day            0
Hit target                       0
Disciplinary failure             0
Education                        0
Son                              0
Social drinker                   0
Social smoker                    0
Pet                              0
Weight                           0
Height                           0
Body mass index                  0
Absenteeism time in hours        0
dtype: int64
```

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

Out[6]:
```
34
```

In [7]:
```
duplicate = data[data.duplicated()]
```

In [8]:
```
duplicate
```

Out[8]:

	ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work	Service time	Age	Work load Average/day	...	Disciplinary failure	Education	Son	Social drinker	s
16	3	23	7	6	1	179	51	18	38	239554	...	0	1	0	1	
68	34	23	10	3	4	118	10	10	37	253465	...	0	1	0	0	
88	28	23	11	4	4	225	26	9	28	306345	...	0	1	1	0	
109	10	22	12	4	4	361	52	3	28	261306	...	0	1	1	1	
127	34	27	1	2	2	118	10	10	37	308593	...	0	1	0	0	
128	34	27	1	3	2	118	10	10	37	308593	...	0	1	0	0	
131	34	27	1	4	2	118	10	10	37	308593	...	0	1	0	0	
132	27	23	1	5	2	184	42	7	27	308593	...	0	1	0	0	
133	34	27	1	5	2	118	10	10	37	308593	...	0	1	0	0	
305	5	23	10	2	4	235	20	13	43	265017	...	0	1	1	1	
351	3	28	12	6	4	179	51	18	38	236629	...	0	1	0	1	
376	3	27	2	4	2	179	51	18	38	251818	...	0	1	0	1	
385	3	27	2	4	2	179	51	18	38	251818	...	0	1	0	1	
386	3	27	2	6	2	179	51	18	38	251818	...	0	1	0	1	
388	3	27	2	4	2	179	51	18	38	251818	...	0	1	0	1	
389	3	27	2	6	2	179	51	18	38	251818	...	0	1	0	1	
440	22	23	5	4	3	179	26	9	30	246074	...	0	3	0	0	

	ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work	Service time	Age	Work load Average/day	...	Disciplinary failure	Education	Son	Social drinker	s
477	24	28	7	3	1	246	25	16	41	230290	...	0	1	0	1	
496	24	28	9	3	1	246	25	16	41	261756	...	0	1	0	1	
560	28	23	12	4	4	225	26	9	28	280549	...	0	1	1	0	
605	3	27	2	4	2	179	51	18	38	264249	...	0	1	0	1	
607	3	27	2	6	2	179	51	18	38	264249	...	0	1	0	1	
610	3	27	2	2	2	179	51	18	38	264249	...	0	1	0	1	
615	3	27	2	4	2	179	51	18	38	264249	...	0	1	0	1	
616	3	27	2	5	2	179	51	18	38	264249	...	0	1	0	1	
631	3	27	3	4	2	179	51	18	38	222196	...	0	1	0	1	
632	3	27	3	5	2	179	51	18	38	222196	...	0	1	0	1	
641	3	27	3	4	2	179	51	18	38	222196	...	0	1	0	1	
643	3	27	3	5	2	179	51	18	38	222196	...	0	1	0	1	
666	22	27	4	6	3	179	26	9	30	246288	...	0	3	0	0	
669	22	27	4	6	3	179	26	9	30	246288	...	0	3	0	0	
673	22	27	4	6	3	179	26	9	30	246288	...	0	3	0	0	
699	15	28	5	5	3	291	31	12	40	237656	...	0	1	1	1	
700	22	27	5	6	3	179	26	9	30	237656	...	0	3	0	0	

34 rows × 21 columns

In [9]:
```
newData = data.drop_duplicates(keep = False, inplace = False)
```

In [10]:
```
newData
```

Out[10]:

	ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work	Service time	Age	Work load Average/day	...	Disciplinary failure	Education	Son	Social drinker	s
0	11	26	7	3	1	289	36	13	33	239554	...	0	1	2	1	
1	36	0	7	3	1	118	13	18	50	239554	...	1	1	1	1	
2	3	23	7	4	1	179	51	18	38	239554	...	0	1	0	1	
3	7	7	7	5	1	279	5	14	39	239554	...	0	1	2	1	
4	11	23	7	5	1	289	36	13	33	239554	...	0	1	2	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
735	11	14	7	3	1	289	36	13	33	264604	...	0	1	2	1	
736	1	11	7	3	1	235	11	14	37	264604	...	0	3	1	0	
737	4	0	0	3	1	118	14	13	40	271219	...	0	1	1	1	
738	8	0	0	4	2	231	35	14	39	271219	...	0	1	2	1	
739	35	0	0	6	3	179	45	14	53	271219	...	0	1	1	0	

680 rows × 21 columns

In [175…

```
cols = ['ID', 'Reason_for_absence', 'Month of absence', 'Day of the week',
       'Seasons', 'Transport expense', 'Distance_from_Residence_to_Work',
       'Service time', 'Age', 'Work load Avg per day', 'Hit target',
       'Disciplinary failure', 'Edu', 'Son', 'Social drinker',
       'Social smoker', 'Pet', 'Wgt', 'Hgt', 'BMI',
       'Absenteeism_time_in_hours']
newData.columns = cols
```

In [176…

```
newData.to_csv('dropped_absenteeism_at_work')
newData.to_csv('/Users/tomisin/Dropbox/My Mac (Tomisins-MacBook-Pro.local)/Documents/MY WORKSPACE/dropped_absenteeism_a
```

In [ ]:

In [177…
```
newData.describe().T
```

Out[177…

	count	mean	std	min	25%	50%	75%	max
ID	680.0	18.260294	10.853892	1.0	10.0	18.0	28.00	36.0
Reason_for_absence	680.0	18.614706	8.520879	0.0	13.0	22.0	25.25	28.0
Month of absence	680.0	6.461765	3.372986	0.0	3.0	6.0	10.00	12.0
Day of the week	680.0	3.876471	1.428525	2.0	3.0	4.0	5.00	6.0
Seasons	680.0	2.551471	1.126857	1.0	2.0	3.0	4.00	4.0
Transport expense	680.0	224.252941	67.402540	118.0	179.0	225.0	260.00	388.0
Distance_from_Residence_to_Work	680.0	29.116176	14.614830	5.0	16.0	26.0	48.00	52.0
Service time	680.0	12.470588	4.362683	1.0	9.0	13.0	16.00	29.0
Age	680.0	36.504412	6.626806	27.0	31.0	37.0	40.00	58.0
Work load Avg per day	680.0	272419.051471	39794.331946	205917.0	244387.0	264604.0	294217.00	378884.0
Hit target	680.0	94.517647	3.822611	81.0	92.0	95.0	97.00	100.0
Disciplinary failure	680.0	0.058824	0.235467	0.0	0.0	0.0	0.00	1.0
Edu	680.0	1.294118	0.672603	1.0	1.0	1.0	1.00	4.0
Son	680.0	1.094118	1.109589	0.0	0.0	1.0	2.00	4.0
Social drinker	680.0	0.564706	0.496160	0.0	0.0	1.0	1.00	1.0
Social smoker	680.0	0.079412	0.270579	0.0	0.0	0.0	0.00	1.0
Pet	680.0	0.785294	1.344083	0.0	0.0	0.0	1.00	8.0
Wgt	680.0	78.994118	12.859040	56.0	69.0	80.0	89.00	108.0
Hgt	680.0	172.275000	6.259905	163.0	169.0	171.0	172.00	196.0
BMI	680.0	26.607353	4.236068	19.0	24.0	25.0	31.00	38.0
Absenteeism_time_in_hours	680.0	7.326471	13.830955	0.0	2.0	3.0	8.00	120.0

In [178…
```
from pandas import read_excel
from numpy import set_printoptions
from sklearn.preprocessing import MinMaxScaler
```

```
array = newData.values

X = array[:,0:20]
Y = array[:,20]
scaler = MinMaxScaler(feature_range=(0, 1))
rescaledX = scaler.fit_transform(X)

print(rescaledX)
```

```
[[0.28571429 0.92857143 0.58333333 ... 0.65384615 0.27272727 0.57894737]
 [1.         0.         0.58333333 ... 0.80769231 0.45454545 0.63157895]
 [0.05714286 0.82142857 0.58333333 ... 0.63461538 0.21212121 0.63157895]
 ...
 [0.08571429 0.         0.         ... 0.80769231 0.21212121 0.78947368]
 [0.2        0.         0.         ... 0.84615385 0.21212121 0.84210526]
 [0.97142857 0.         0.         ... 0.40384615 0.36363636 0.31578947]]
```

In [180…

```
rescaledXDF = pd.DataFrame(rescaledX, columns =['ID', 'Reason for absence', 'Month of absence', 'Day of the week',
       'Seasons', 'Transportation expense', 'Distance from Residence to Work',
       'Service time', 'Age', 'Work load Average/day ', 'Hit target',
       'Disciplinary failure', 'Education', 'Son', 'Social drinker',
       'Social smoker', 'Pet', 'Weight', 'Height', 'Body mass index'])
```

In [181…

```
rescaledXDF
```

Out[181…

	ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work	Service time	Age	Work load Average/day	Hit target	Disciplinary failure	Educ
0	0.285714	0.928571	0.583333	0.25	0.000000	0.633333	0.659574	0.428571	0.193548	0.194471	0.842105	0.0	0.00
1	1.000000	0.000000	0.583333	0.25	0.000000	0.000000	0.170213	0.607143	0.741935	0.194471	0.842105	1.0	0.00
2	0.057143	0.821429	0.583333	0.50	0.000000	0.225926	0.978723	0.607143	0.354839	0.194471	0.842105	0.0	0.00
3	0.171429	0.250000	0.583333	0.75	0.000000	0.596296	0.000000	0.464286	0.387097	0.194471	0.842105	0.0	0.00
4	0.285714	0.821429	0.583333	0.75	0.000000	0.633333	0.659574	0.428571	0.193548	0.194471	0.842105	0.0	0.00
…	…	…	…	…	…	…	…	…	…	…	…	…	
675	0.285714	0.500000	0.583333	0.25	0.000000	0.633333	0.659574	0.428571	0.193548	0.339296	0.631579	0.0	0.00

	ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work	Service time	Age	Work load Average/day	Hit target	Disciplinary failure	Educ
676	0.000000	0.392857	0.583333	0.25	0.000000	0.433333	0.127660	0.464286	0.322581	0.339296	0.631579	0.0	0.66
677	0.085714	0.000000	0.000000	0.25	0.000000	0.000000	0.191489	0.428571	0.419355	0.377540	0.736842	0.0	0.00
678	0.200000	0.000000	0.000000	0.50	0.333333	0.418519	0.638298	0.464286	0.387097	0.377540	0.736842	0.0	0.00
679	0.971429	0.000000	0.000000	1.00	0.666667	0.225926	0.851064	0.464286	0.838710	0.377540	0.736842	0.0	0.00

680 rows × 20 columns

In [182…

```
rescaledXDF.describe().T
```

Out[182…

	count	mean	std	min	25%	50%	75%	max
ID	680.0	0.493151	0.310111	0.0	0.257143	0.485714	0.771429	1.0
Reason for absence	680.0	0.664811	0.304317	0.0	0.464286	0.785714	0.901786	1.0
Month of absence	680.0	0.538480	0.281082	0.0	0.250000	0.500000	0.833333	1.0
Day of the week	680.0	0.469118	0.357131	0.0	0.250000	0.500000	0.750000	1.0
Seasons	680.0	0.517157	0.375619	0.0	0.333333	0.666667	1.000000	1.0
Transportation expense	680.0	0.393529	0.249639	0.0	0.225926	0.396296	0.525926	1.0
Distance from Residence to Work	680.0	0.513110	0.310954	0.0	0.234043	0.446809	0.914894	1.0
Service time	680.0	0.409664	0.155810	0.0	0.285714	0.428571	0.535714	1.0
Age	680.0	0.306594	0.213768	0.0	0.129032	0.322581	0.419355	1.0
Work load Average/day	680.0	0.384478	0.230069	0.0	0.222412	0.339296	0.510502	1.0
Hit target	680.0	0.711455	0.201190	0.0	0.578947	0.736842	0.842105	1.0
Disciplinary failure	680.0	0.058824	0.235467	0.0	0.000000	0.000000	0.000000	1.0
Education	680.0	0.098039	0.224201	0.0	0.000000	0.000000	0.000000	1.0
Son	680.0	0.273529	0.277397	0.0	0.000000	0.250000	0.500000	1.0

	count	mean	std	min	25%	50%	75%	max
Social drinker	680.0	0.564706	0.496160	0.0	0.000000	1.000000	1.000000	1.0
Social smoker	680.0	0.079412	0.270579	0.0	0.000000	0.000000	0.000000	1.0
Pet	680.0	0.098162	0.168010	0.0	0.000000	0.000000	0.125000	1.0
Weight	680.0	0.442195	0.247289	0.0	0.250000	0.461538	0.634615	1.0
Height	680.0	0.281061	0.189694	0.0	0.181818	0.242424	0.272727	1.0
Body mass index	680.0	0.400387	0.222951	0.0	0.263158	0.315789	0.631579	1.0

In [183…

```
from sklearn.preprocessing import Normalizer
from pandas import read_excel
from numpy import set_printoptions

array = rescaledXDF.values

X = array[:,0:20]

scaler = Normalizer().fit(X)
normalizedX = scaler.transform(X)
```

In [184…

```
normalizedXDF = pd.DataFrame(normalizedX, columns =['ID', 'Reason for absence', 'Month of absence', 'Day of the week',
       'Seasons', 'Transportation expense', 'Distance from Residence to Work',
       'Service time', 'Age', 'Work load Average/day ', 'Hit target',
       'Disciplinary failure', 'Education', 'Son', 'Social drinker',
       'Social smoker', 'Pet', 'Weight', 'Height', 'Body mass index'])
```

In [185…

```
normalizedXDF
```

Out[185…

	ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work	Service time	Age	Work load Average/day	Hit target	Disciplinary failure	E
0	0.124653	0.405123	0.254500	0.109072	0.000000	0.276315	0.287763	0.186980	0.084442	0.084845	0.367399	0.000000	
1	0.394725	0.000000	0.230256	0.098681	0.000000	0.000000	0.067187	0.239654	0.292860	0.076762	0.332400	0.394725	

	ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work	Service time	Age	Work load Average/day	Hit target	Disciplinary failure	E
2	0.024670	0.354626	0.251836	0.215859	0.000000	0.097536	0.422533	0.262115	0.153191	0.083957	0.363553	0.000000	
3	0.077777	0.113425	0.264658	0.340275	0.000000	0.270539	0.000000	0.210646	0.175626	0.088231	0.382063	0.000000	
4	0.121103	0.348172	0.247253	0.317897	0.000000	0.268446	0.279569	0.181655	0.082038	0.082429	0.356936	0.000000	
...	...	...	...	...	...	...	...	...	...	...	...	...	
675	0.136079	0.238138	0.277828	0.119069	0.000000	0.301642	0.314140	0.204118	0.092183	0.161599	0.300806	0.000000	
676	0.000000	0.232811	0.345690	0.148153	0.000000	0.256798	0.075652	0.275141	0.191165	0.201071	0.374281	0.000000	
677	0.040251	0.000000	0.000000	0.117400	0.000000	0.000000	0.089923	0.201257	0.196929	0.177293	0.346021	0.000000	
678	0.091124	0.000000	0.000000	0.227810	0.151874	0.190686	0.290822	0.211538	0.176369	0.172015	0.335720	0.000000	
679	0.424326	0.000000	0.000000	0.436806	0.291204	0.098686	0.371750	0.202803	0.366353	0.164912	0.321857	0.000000	

680 rows × 20 columns

In [186…

```
normalizedXDF.describe().T
```

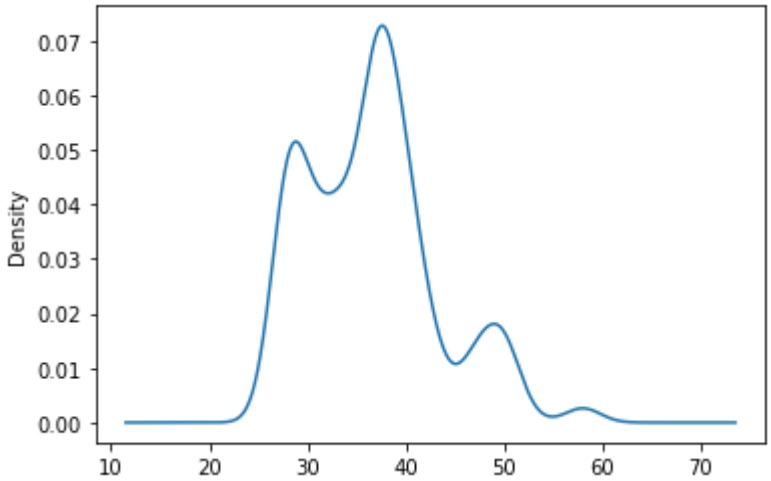
Out[186…

	count	mean	std	min	25%	50%	75%	max
ID	680.0	0.225946	0.149393	0.0	0.111922	0.208612	0.358579	0.574804
Reason for absence	680.0	0.297200	0.135681	0.0	0.203181	0.332037	0.394570	0.539647
Month of absence	680.0	0.237660	0.120347	0.0	0.126684	0.246425	0.340648	0.524836
Day of the week	680.0	0.204121	0.151642	0.0	0.100017	0.206045	0.340850	0.523705
Seasons	680.0	0.224650	0.159684	0.0	0.128035	0.251813	0.366560	0.563160
Transportation expense	680.0	0.175373	0.109179	0.0	0.095223	0.187884	0.236702	0.485164
Distance from Residence to Work	680.0	0.224778	0.126676	0.0	0.109245	0.218856	0.335105	0.485233
Service time	680.0	0.181415	0.062989	0.0	0.145705	0.178944	0.230232	0.370569
Age	680.0	0.133428	0.092865	0.0	0.056812	0.136240	0.177907	0.593706

	count	mean	std	min	25%	50%	75%	max
Work load Average/day	680.0	0.171939	0.102784	0.0	0.096918	0.147702	0.220104	0.529239
Hit target	680.0	0.321565	0.102932	0.0	0.260117	0.324646	0.384404	0.603193
Disciplinary failure	680.0	0.023633	0.095029	0.0	0.000000	0.000000	0.000000	0.510167
Education	680.0	0.049219	0.113376	0.0	0.000000	0.000000	0.000000	0.425600
Son	680.0	0.117842	0.114744	0.0	0.000000	0.106125	0.202270	0.452420
Social drinker	680.0	0.234744	0.207836	0.0	0.000000	0.373720	0.420134	0.543198
Social smoker	680.0	0.034115	0.116877	0.0	0.000000	0.000000	0.000000	0.525699
Pet	680.0	0.044303	0.073327	0.0	0.000000	0.000000	0.058423	0.469600
Weight	680.0	0.193320	0.102545	0.0	0.108907	0.204530	0.277301	0.420958
Height	680.0	0.125665	0.081905	0.0	0.084926	0.103909	0.137663	0.460974
Body mass index	680.0	0.175374	0.092290	0.0	0.113076	0.156700	0.252484	0.437797

In [190…
```
newData.Age.plot(kind='density')
```

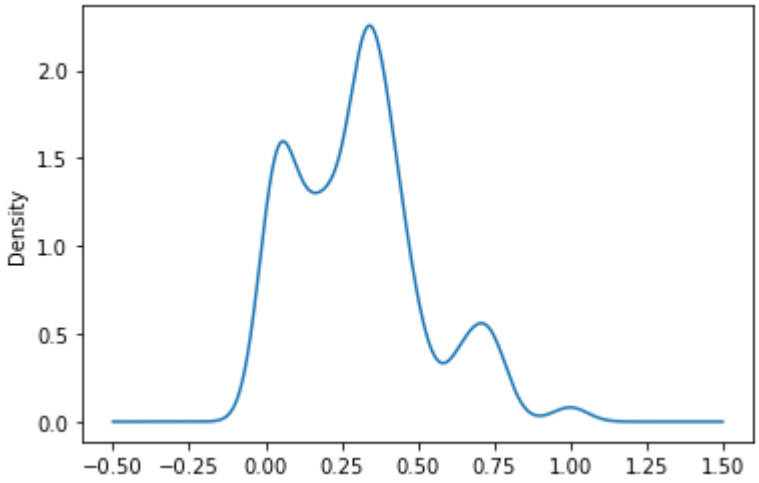
Out[190…
```
<AxesSubplot:ylabel='Density'>
```



In [189…
```
rescaledXDF.Age.plot(kind='density')
```

Out[189…

```
<AxesSubplot:ylabel='Density'>
```

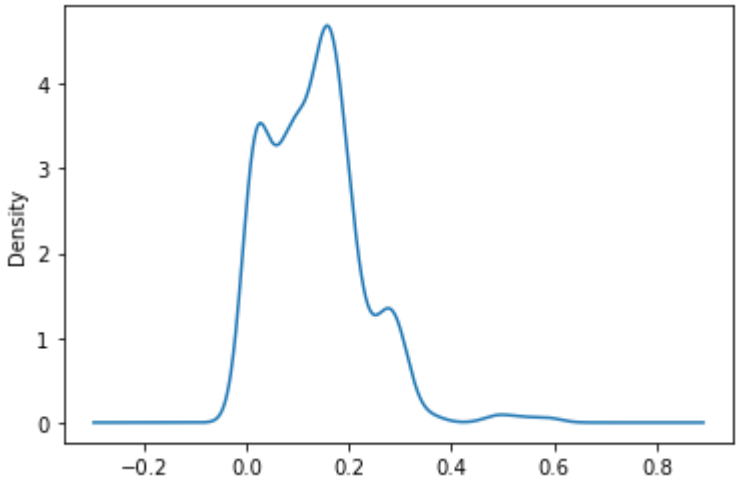


In [188…

```
normalizedXDF.Age.plot(kind='density')
```

Out[188…

```
<AxesSubplot:ylabel='Density'>
```



In [222…

```
normalizedXDF['Absenteeism_time_in_hours'] = newData['Absenteeism_time_in_hours']
```

In [223…
```
normalizedXDF.to_csv('cleaned_absenteeism_at_work')
```

In [224…
```
normalizedXDF.to_csv('/Users/tomisin/Dropbox/My Mac (Tomisins-MacBook-Pro.local)/Documents/MY WORKSPACE/cleaned_absente
```

In [225…
```
''' 2. Implement univariate analysis of the features in the data and surface key insights '''.
'''Here, we only need to consider labels that have a noticeable trend,
whether based on logical/human reasoning or not
'''

***''''ID'- Their ID's do not have a relationship with their absenteeism from work.
There is no noticeable trend at all.

***'Reason_for_absence'- Employees have varying reasons for being absent
ranging from the more common reason to the least.
I think this label will not be useful for the model.

***'Month of absence'- The sb countplot shows that employees are generally absent from work all year round.
But we observe gradual increase in absence in some consecutive months
during the summer period or winter period for example where people are expected to go on vacations
In each quarter of the year, absenteeism either increases or decreases.
There is no particular month they are peculiarly absent from work due to whatever reason.
I think this label will be useful for our output prediction in our model. Hence, it should not be incorporated.


***'Day of the week'- It appears that there is usually much work load on monday
which progressively reduces into the week (looks like a trend). On fridays again,
there is a lot of work to be done and employees purposely cut work as there is no fixed working days or times.
I think this label will be useful for our output prediction in our model


***'Seasons'- There are 4 seasons and employees are absent all season round.
There seams to be a trend wherein  absenteeism increases in colder seasons.
This may be due to harsh weather conditions during this periods.
Some people get sick when there is a slight change in weather conditions.
We can therefore not see a trend in their absenteeism.
I think this label will be useful for our output prediction in our model

***'Transport expense'- Their transport expenses have nothing to do with their absenteeism fromm work.
There is no trend at all. I do not think this label or features in this label will be useful for the model.

***'Distance_from_Residence_to_Work'- Their distances from residence to work
```

have no relatiosnship to their absenteeism. For this reason,
I do not think this label or features in this label will be useful for the model.

***'Service time'- Their service times have no relationship to their absenteeism from work.
It appears that employees don't have a preferred/fixed working time.
This label will not be useful in determining their absent times in hours

***'Age'- Their ages are have no relationship to their absenteeism from work.
This label will not be useful in determining their absent times in hours

***'Work load Avg/day'- Employees do not have equal amount of work load.
This could be a cause of absenteeism at work.
 but because their absenteeism is inversely proportional
to their work load average per day, I see no trend at all.
This label will not be useful in determining their absent times in hours.

***'Hit target',- Their hit targets have no relationship to their absenteeism from work.
There is no obvious trend with their absenteeism from work
This label will not be useful in determining their absent times in hours

***'Disciplinary failure'- I cannot comprehend this at all.

***'Edu'- There is a relationship between absenteeism in young adults than in older ones.
Younger adults are probably more irresponsible, lazier/less focused
than older ones (graduates and postgraduates)
or maybe due to the fact that they have less work experience
than the graduates or for whatever reasons I can't think of right now.

***'Son'- Here, we can see that the number of children possessed has an influence on their absenteeism.
It appears that employees with the least number of children are more absent from work for whatever reasons.

***'Social drinker'- Here, about 380 people are social drinkers and 300 non-drinkers.
The employees in these two groups are absent to almost the same degree.
This label will not be useful in our model output prediction.

***'Social smoker'- Here, about 630 people are social non-smokers and 50 social smokers.
The employees in these two groups are absent to varying degree.
This label can be useful in our model output prediction.

***'Pet'- There is a relationship in absenteeism between those who have no pets
and those who have more than 1 pet. We see a trend and observe more absence in those
who have no pets than in those who have 1 or 2 or 4. This label can be useful in our model output prediction.

```
***'Wgt'- Their weights have no influence on their absenteeism.
This label will not be useful in our model output prediction.


***'Hgt'- Their heights have no influence on their absenteeism.
This label will not be useful in our model output prediction

'''
```

```
  File "/var/folders/j0/1q4cz_pd0631rv8ljk2flvqr0000gn/T/ipykernel_22611/2461354783.py", line 1
    ''' 2. Implement univariate analysis of the features in the data and surface key insights '''.
                                                                                                 ^
SyntaxError: invalid syntax
```

In [ ]:

In [196…

```
import pandas as pd
fileName = '/Users/tomisin/Dropbox/My Mac (Tomisins-MacBook-Pro.local)/Documents/MY WORKSPACE/dropped_absenteeism_at_wo

df = read_csv(fileName)
```

In [219…

```
df
```

Out[219…

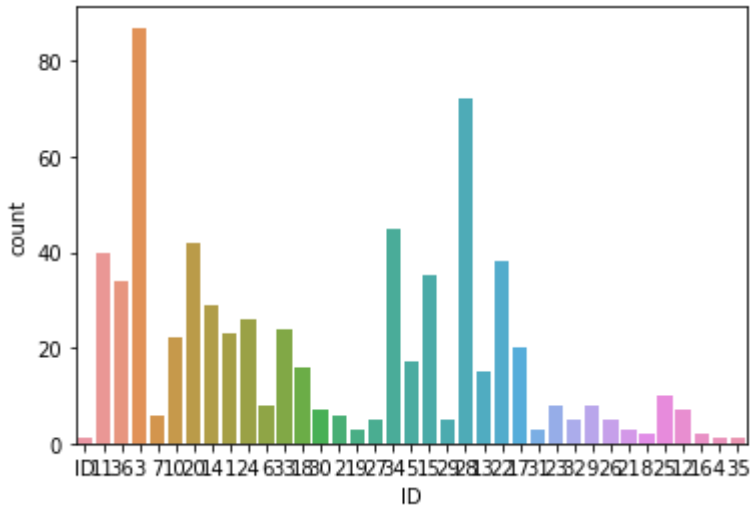
	Unnamed: 0	ID	Reason_for_absence	Month of absence	Day of the week	Seasons	Transport expense	Distance_from_Residence_to_Work	Service time	Age	...	Disciplinary failure
0	0	11	26	7	3	1	289	36	13	33	...	0
1	1	36	0	7	3	1	118	13	18	50	...	1
2	2	3	23	7	4	1	179	51	18	38	...	0
3	3	7	7	7	5	1	279	5	14	39	...	0
4	4	11	23	7	5	1	289	36	13	33	...	0
...	...	...	...	...	...	...	...	...	...	...	...	...
675	735	11	14	7	3	1	289	36	13	33	...	0

	Unnamed: 0	ID	Reason_for_absence	Month of absence	Day of the week	Seasons	Transport expense	Distance_from_Residence_to_Work	Service time	Age	...	Disciplinary failure
676	736	1	11	7	3	1	235	11	14	37	...	0
677	737	4	0	0	3	1	118	14	13	40	...	0
678	738	8	0	0	4	2	231	35	14	39	...	0
679	739	35	0	0	6	3	179	45	14	53	...	0

680 rows × 22 columns

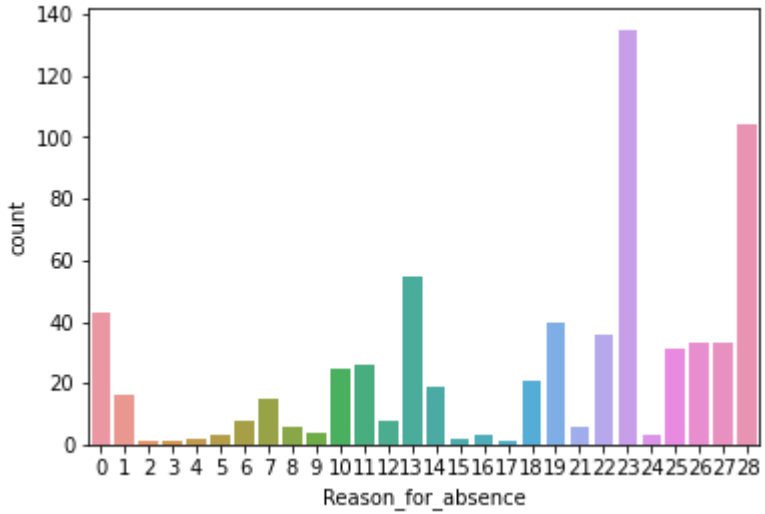
In [226…
```
sb.countplot(x = 'ID', data = df)
```

Out[226…
```
<AxesSubplot:xlabel='ID', ylabel='count'>
```



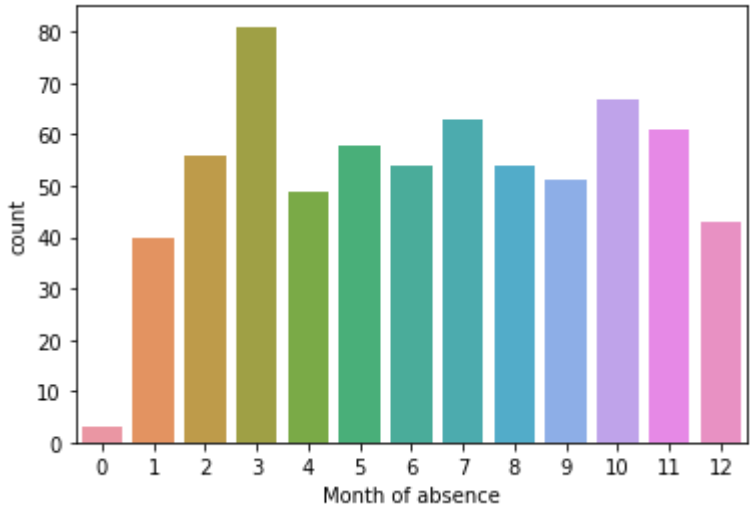
In [198…
```
sb.countplot(x = 'Reason_for_absence', data = df)
```

Out[198…
```
<AxesSubplot:xlabel='Reason_for_absence', ylabel='count'>
```



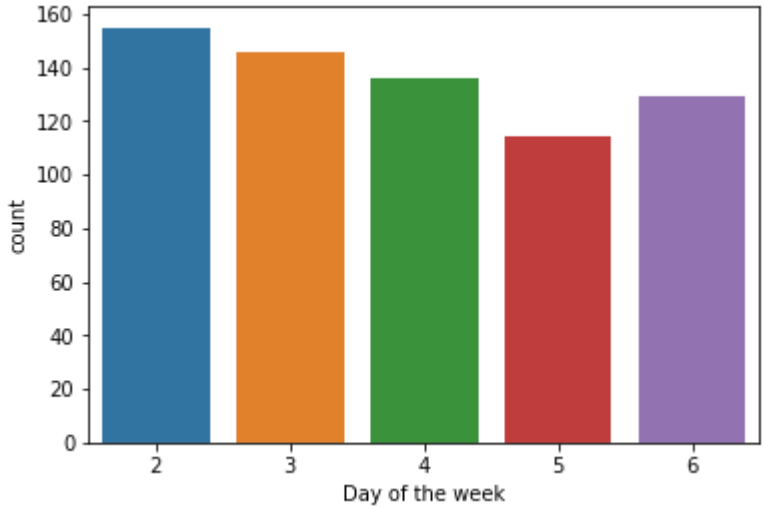
In [199…
```
sb.countplot(x ='Month of absence', data = df)
```

Out[199…
```
<AxesSubplot:xlabel='Month of absence', ylabel='count'>
```



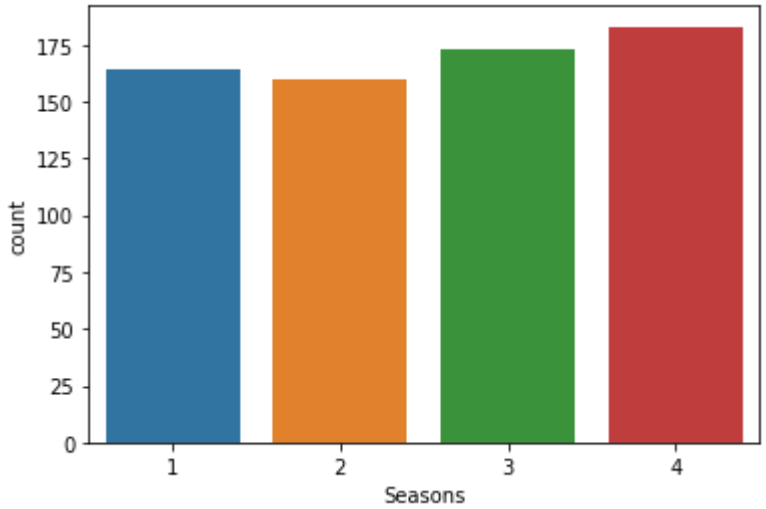
In [200…
```
sb.countplot(x ='Day of the week', data = df)
```

Out[200…
```
<AxesSubplot:xlabel='Day of the week', ylabel='count'>
```



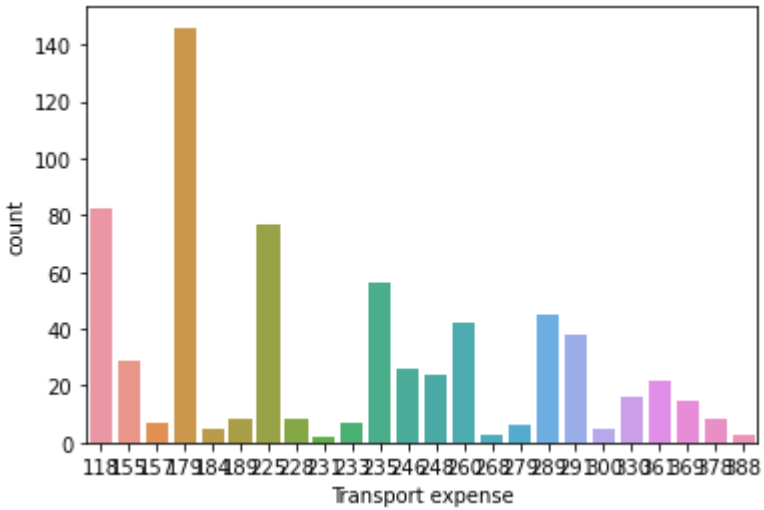
In [201…
```
sb.countplot(x ='Seasons', data = df)
```

Out[201…
```
<AxesSubplot:xlabel='Seasons', ylabel='count'>
```



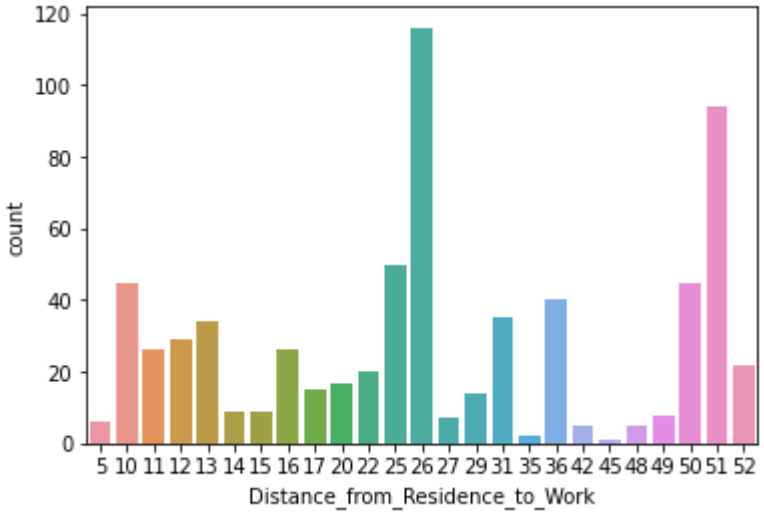
In [202…
```
sb.countplot(x ='Transport expense', data = df)
```

Out[202…
```
<AxesSubplot:xlabel='Transport expense', ylabel='count'>
```



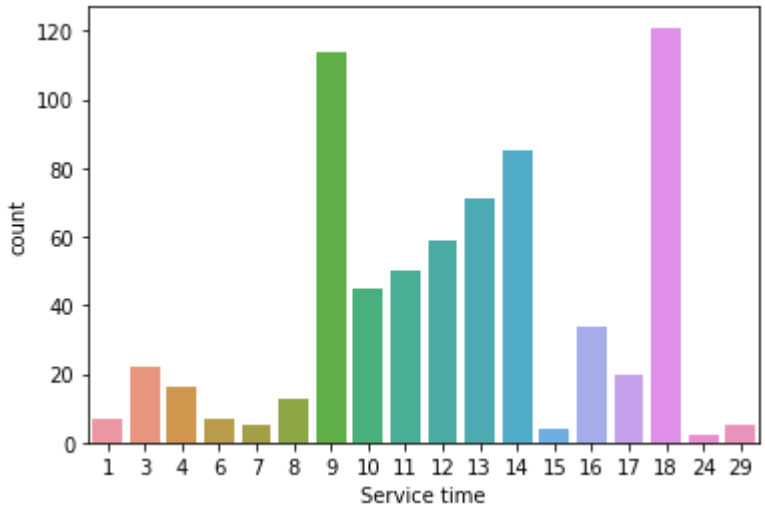
In [203…
```
sb.countplot(x ='Distance_from_Residence_to_Work', data = df)
```

Out[203…
```
<AxesSubplot:xlabel='Distance_from_Residence_to_Work', ylabel='count'>
```



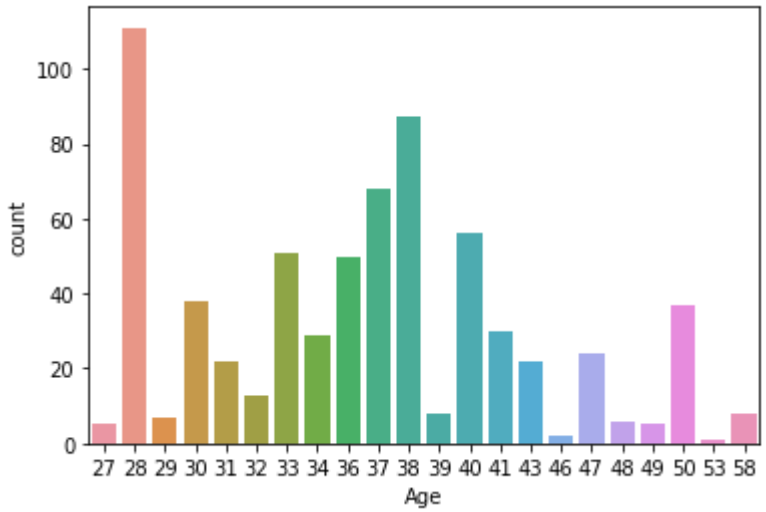
In [204…
```
sb.countplot(x ='Service time', data = df)
```

Out[204…
```
<AxesSubplot:xlabel='Service time', ylabel='count'>
```



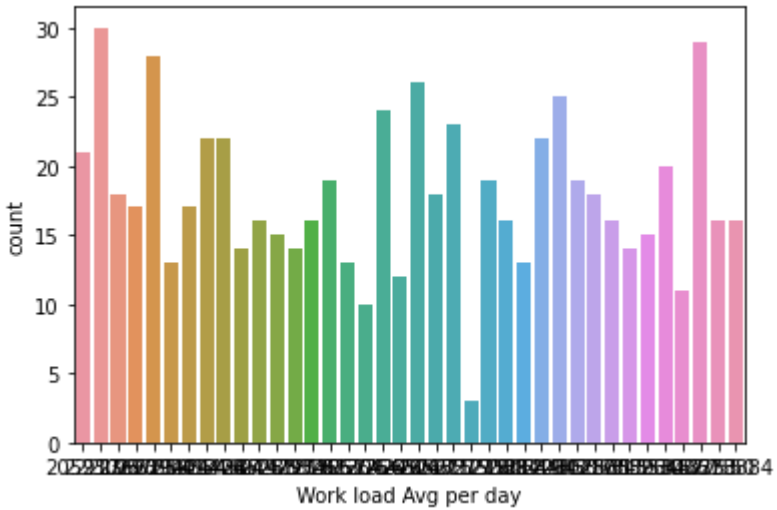
In [205…
```
sb.countplot(x ='Age', data = df)
```

Out[205…
```
<AxesSubplot:xlabel='Age', ylabel='count'>
```



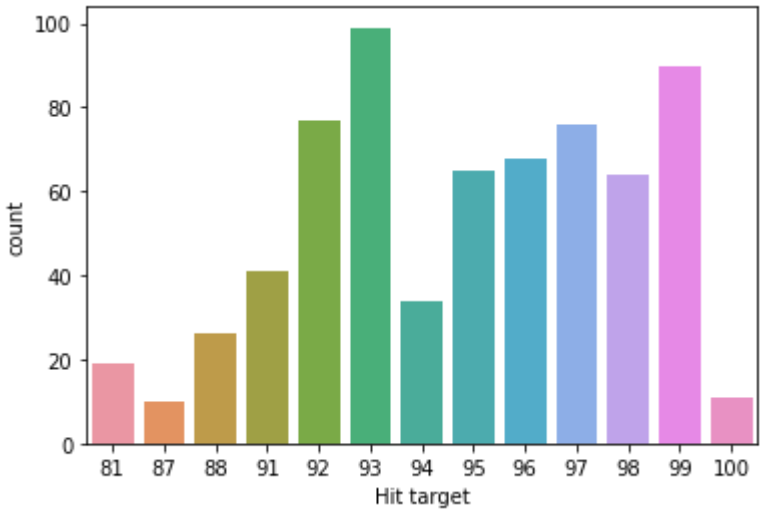
In [206…
```
sb.countplot(x ='Work load Avg per day', data = df)
```

Out[206…
```
<AxesSubplot:xlabel='Work load Avg per day', ylabel='count'>
```



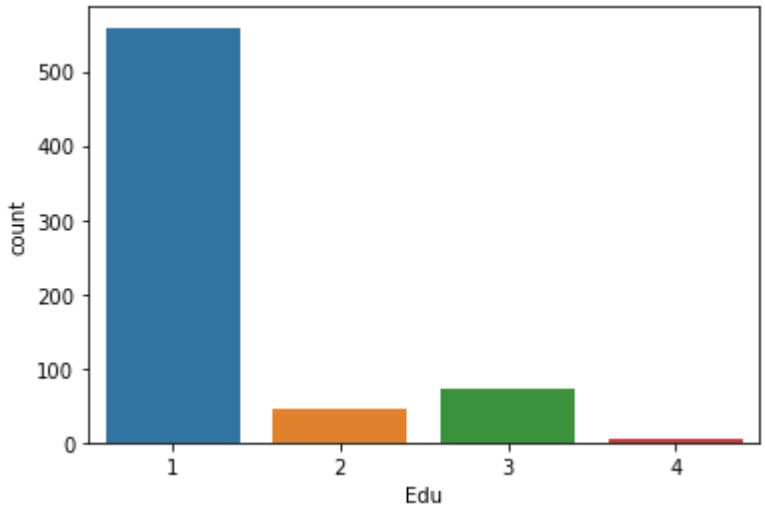
In [207…
```
sb.countplot(x ='Hit target', data = df)
```

Out[207…
```
<AxesSubplot:xlabel='Hit target', ylabel='count'>
```



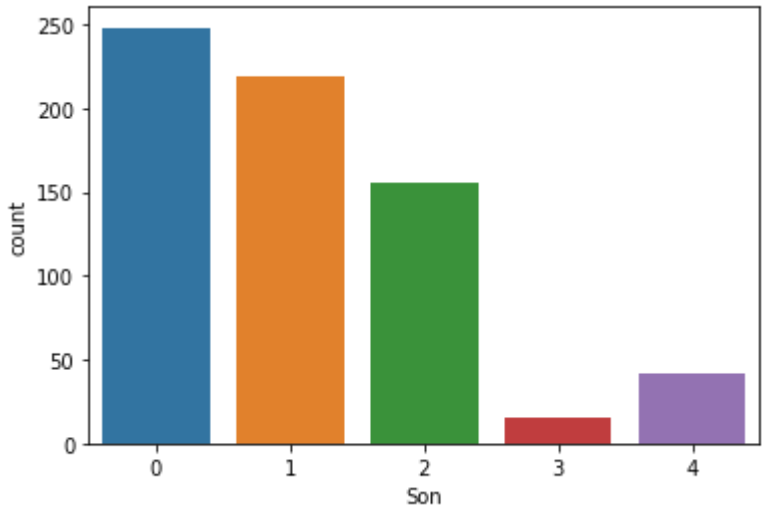
In [208…
```
sb.countplot(x ='Edu', data = df)
```

Out[208…
```
<AxesSubplot:xlabel='Edu', ylabel='count'>
```



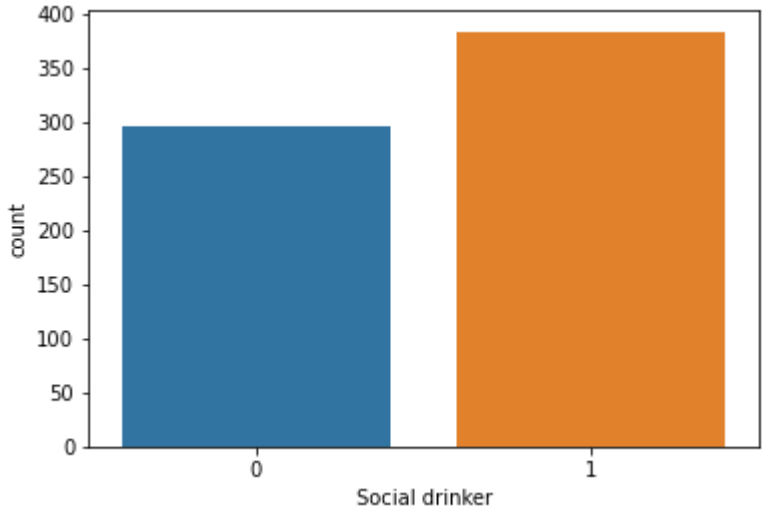
In [209…
```
sb.countplot(x ='Son', data = df)
```

Out[209…
```
<AxesSubplot:xlabel='Son', ylabel='count'>
```



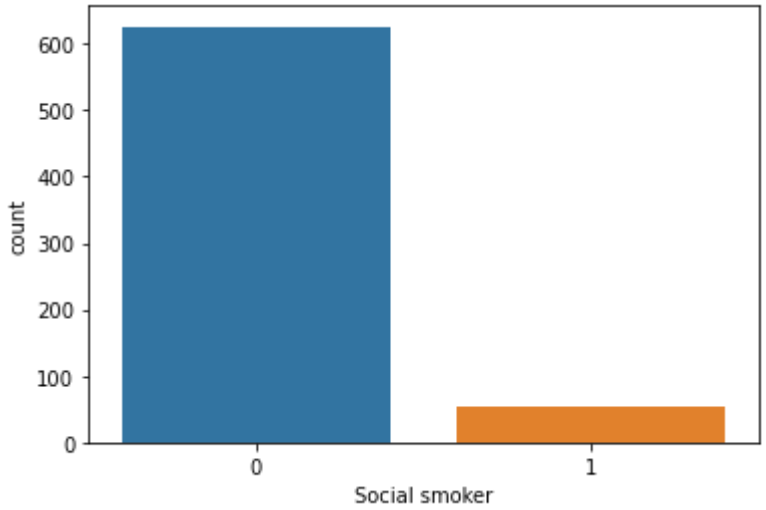
In [210…
```
sb.countplot(x ='Social drinker', data = df)
```

Out[210…
```
<AxesSubplot:xlabel='Social drinker', ylabel='count'>
```



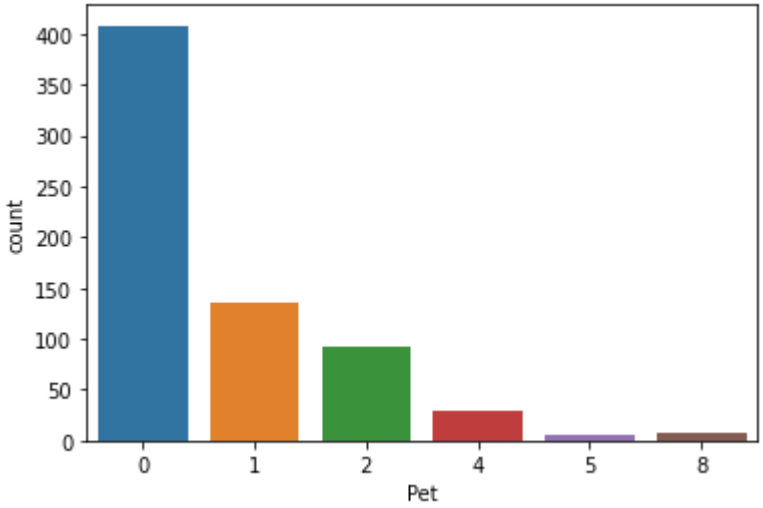
In [211…
```
sb.countplot(x ='Social smoker', data = df)
```

Out[211…
```
<AxesSubplot:xlabel='Social smoker', ylabel='count'>
```



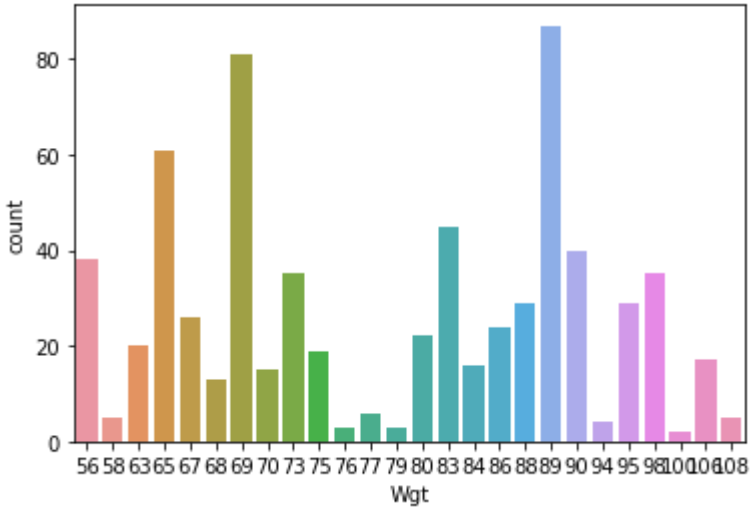
In [212…
```
sb.countplot(x ='Pet', data = df)
```

Out[212…
```
<AxesSubplot:xlabel='Pet', ylabel='count'>
```



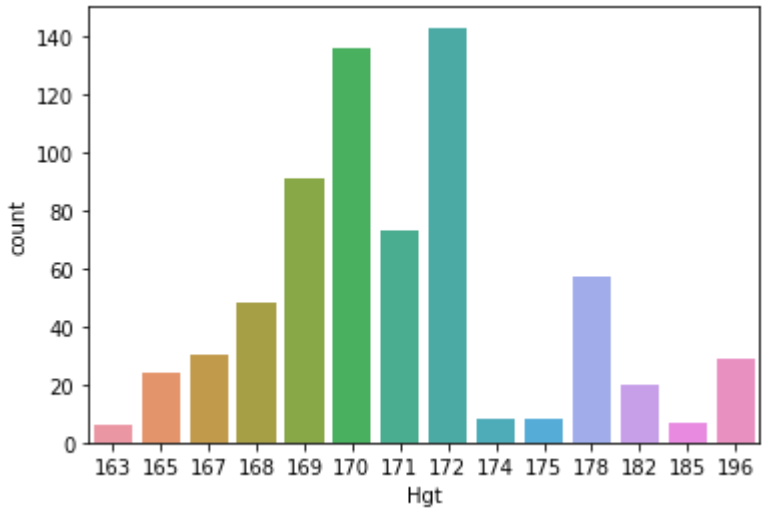
In [213…
```
sb.countplot(x ='Wgt', data = df)
```

Out[213…
```
<AxesSubplot:xlabel='Wgt', ylabel='count'>
```



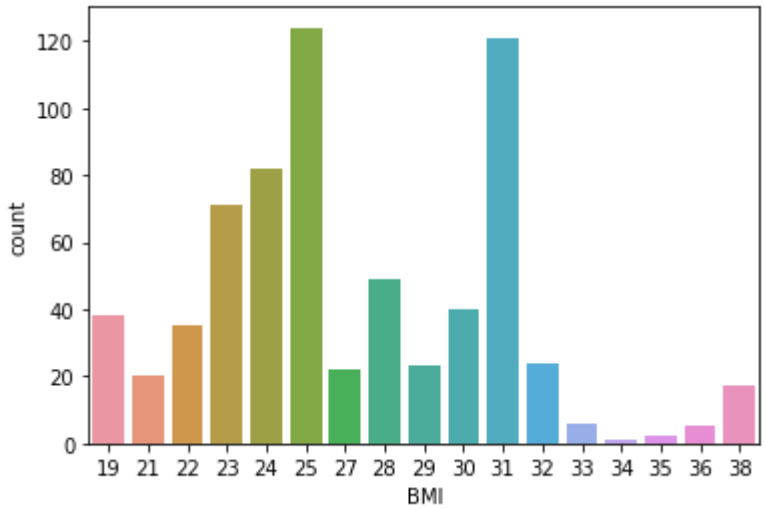
In [214…
```
sb.countplot(x ='Hgt', data = df)
```

Out[214…
```
<AxesSubplot:xlabel='Hgt', ylabel='count'>
```



In [215…
```
sb.countplot(x ='BMI', data = df)
```

Out[215…
```
<AxesSubplot:xlabel='BMI', ylabel='count'>
```



In [ ]:

In [216…
```
df.dtypes
```

Out[216…

```
Unnamed: 0                        int64
ID                                int64
Reason_for_absence                int64
Month of absence                  int64
Day of the week                   int64
Seasons                           int64
Transport expense                 int64
Distance_from_Residence_to_Work   int64
Service time                      int64
Age                               int64
Work load Avg per day             int64
Hit target                        int64
Disciplinary failure              int64
Edu                               int64
Son                               int64
Social drinker                    int64
Social smoker                     int64
Pet                               int64
Wgt                               int64
Hgt                               int64
BMI                               int64
Absenteeism_time_in_hours         int64
dtype: object
```

In [217…

```
df.describe().T
```

Out[217…

	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	680.0	366.302941	212.332254	0.0	187.75	361.5	544.25	739.0
ID	680.0	18.260294	10.853892	1.0	10.00	18.0	28.00	36.0
Reason_for_absence	680.0	18.614706	8.520879	0.0	13.00	22.0	25.25	28.0
Month of absence	680.0	6.461765	3.372986	0.0	3.00	6.0	10.00	12.0
Day of the week	680.0	3.876471	1.428525	2.0	3.00	4.0	5.00	6.0
Seasons	680.0	2.551471	1.126857	1.0	2.00	3.0	4.00	4.0
Transport expense	680.0	224.252941	67.402540	118.0	179.00	225.0	260.00	388.0
Distance_from_Residence_to_Work	680.0	29.116176	14.614830	5.0	16.00	26.0	48.00	52.0
Service time	680.0	12.470588	4.362683	1.0	9.00	13.0	16.00	29.0

	count	mean	std	min	25%	50%	75%	max
Age	680.0	36.504412	6.626806	27.0	31.00	37.0	40.00	58.0
Work load Avg per day	680.0	272419.051471	39794.331946	205917.0	244387.00	264604.0	294217.00	378884.0
Hit target	680.0	94.517647	3.822611	81.0	92.00	95.0	97.00	100.0
Disciplinary failure	680.0	0.058824	0.235467	0.0	0.00	0.0	0.00	1.0
Edu	680.0	1.294118	0.672603	1.0	1.00	1.0	1.00	4.0
Son	680.0	1.094118	1.109589	0.0	0.00	1.0	2.00	4.0
Social drinker	680.0	0.564706	0.496160	0.0	0.00	1.0	1.00	1.0
Social smoker	680.0	0.079412	0.270579	0.0	0.00	0.0	0.00	1.0
Pet	680.0	0.785294	1.344083	0.0	0.00	0.0	1.00	8.0
Wgt	680.0	78.994118	12.859040	56.0	69.00	80.0	89.00	108.0
Hgt	680.0	172.275000	6.259905	163.0	169.00	171.0	172.00	196.0
BMI	680.0	26.607353	4.236068	19.0	24.00	25.0	31.00	38.0
Absenteeism_time_in_hours	680.0	7.326471	13.830955	0.0	2.00	3.0	8.00	120.0

In [220…
```
df.columns
```

Out[220…
```
Index(['Unnamed: 0', 'ID', 'Reason_for_absence', 'Month of absence',
       'Day of the week', 'Seasons', 'Transport expense',
       'Distance_from_Residence_to_Work', 'Service time', 'Age',
       'Work load Avg per day', 'Hit target', 'Disciplinary failure', 'Edu',
       'Son', 'Social drinker', 'Social smoker', 'Pet', 'Wgt', 'Hgt', 'BMI',
       'Absenteeism_time_in_hours'],
      dtype='object')
```

In [ ]:
```
```

In [171…
```
'''3. On the basis of ONLY univariate analytics, describe the key drivers of absenteeism in the data

Based oon explanations given in question 2 above, The following labels are my most preferred (key drivers) of absenteei
```

```
2. Edu
3. Son
4. Social smoker
5. Pet
6. Month of absence
7. Day of the week

'''
```

Out[171… '3. On the basis of ONLY univariate analytics, describe the key drivers of absenteeism in the data'

In [ ]:

In [ ]: