MACHINE LEARNING AND DATA SCIENCE

CAPE STONE PROJECT

Dataset:

teen_phone_addition_dataset

Domain:

Supervised Learning.(Requirements clear. Proper Input & Output.)

The output of our dataset is Continuous Value. So it comes under Regression.

Finding Best Model:

Using **Feature Selection Method (selectkbest)** we can find out the best model. Using this method, we identified **SUPPORT VECTOR MACHINE NON LINEAR** is the best model. So we forward the model in deployment phase.

	Linear	SVMl	SVMnl	Decision	Random
f_regression	0.722553	0.712625	0.985651	0.690482	0.886317

Best Feature in our dataset:

```
Top 5 selected features:
Index(['Daily_Usage_Hours', 'Phone_Checks_Per_Day', 'Apps_Used_Daily',
       'Time_on_Social_Media', 'Time_on_Gaming'],
      dtype='object')
```

During deployment phase, we should only assign value to the column name mentioned above.

Before deployment phase we should

- Check NULL Values,Missing Values
- Do Data Preprocessing
- Find outliers

Check NULL Values,Missing Values:

Using the following coding we can check null values

dataset.isnull().sum()

```
ID                        0
Name                      0
Age                       0
Gender                    0
Location                  0
School_Grade              0
Daily_Usage_Hours         0
Sleep_Hours               0
Academic_Performance      0
Social_Interactions       0
Exercise_Hours            0
Anxiety_Level             0
Depression_Level          0
Self_Esteem               0
Parental_Control          0
Screen_Time_Before_Bed    0
Phone_Checks_Per_Day      0
Apps_Used_Daily           0
Time_on_Social_Media      0
Time_on_Gaming            0
Time_on_Education         0
Phone_Usage_Purpose       0
Family_Communication      0
Weekend_Usage_Hours       0
Addiction_Level           0
dtype: int64
```

Data Science-Univariant Analysis

Finding outlier:

- If Min value is less than Lesser Outlier then the particular column is Lesser Outlier.
- If Max value is greater than Greater Outlier then the particular column is Greater Outlier

In our dataset Outliers are listed below

```
Lesser: ['Weekend_Usage_Hours', 'Addiction_Level']
Greater: ['Daily_Usage_Hours', 'Exercise_Hours', 'Screen_Time_Before_Bed', 'Weekend_Usage_Hours']
```

Handling Outlier:

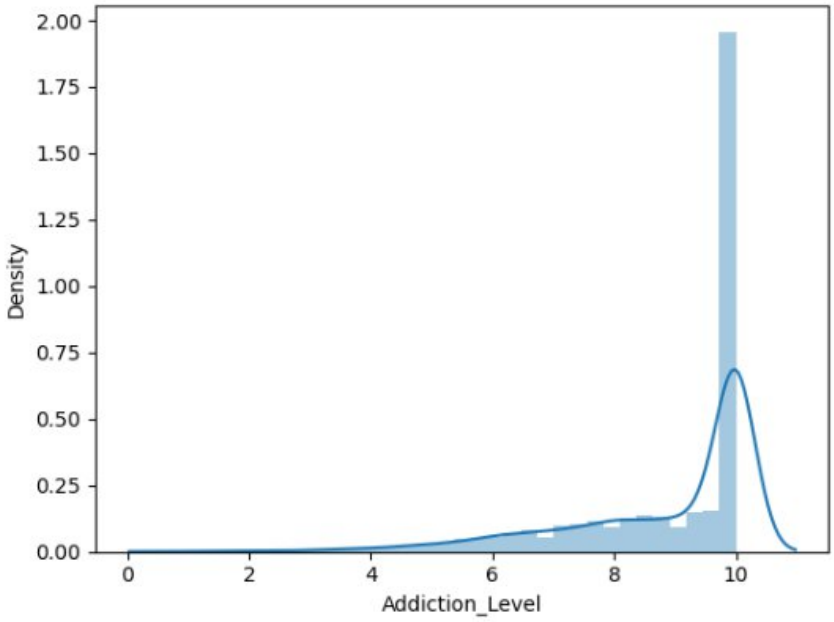
Lesser Outlier is replace with MIN Value.Greater Outlier is replace with MAX Value.

```
Lesser, Greater = Finding_outliers(descriptive, quan)
print("Lesser:", Lesser)
print("Greater:", Greater)
```

```
Lesser: []
Greater: []
```

Normal Distribution:

A large number of teens have low addiction level.Moverover smaller number of teens have high addictional level that is shown in following Fig(1).



Fig(1)

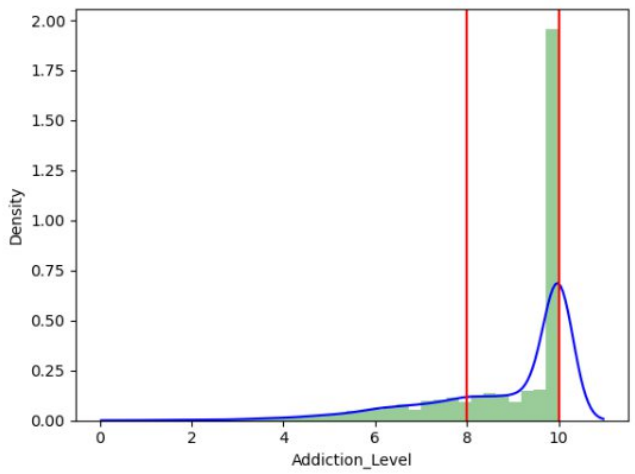
Probability Density Function:

```
get_pdf_prob(dataset["Addiction_Level"],8,10)
```

```
Mean-8.882,Standard Deviation-1.610
The area between range(8,10):0.46049331442644087
```

Approximately **46.04%,** that is nearly half of teens in the dataset fall with in this range (8 to 10).See the following Fig(2)

]: 0.46049331442644087

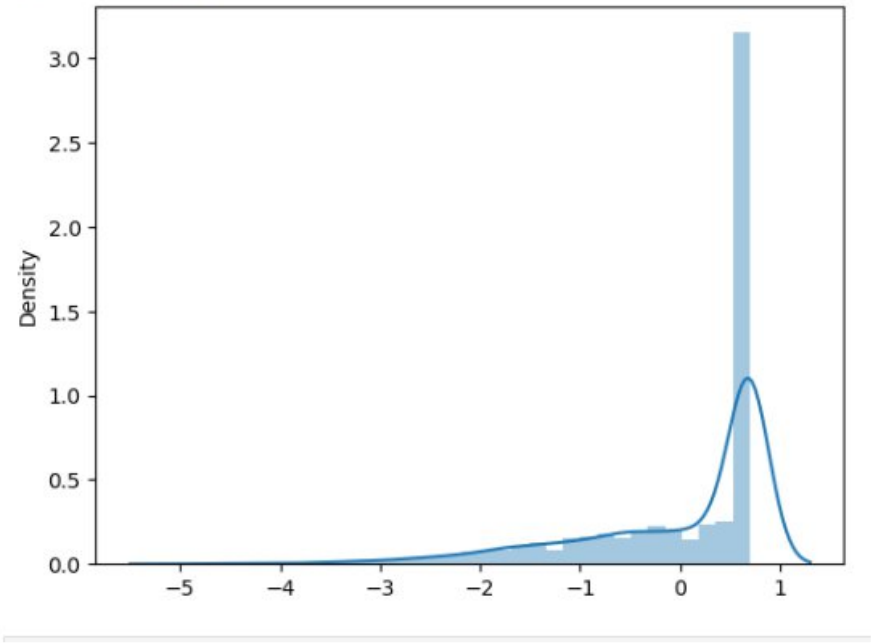


Fig(2)

Standard Normal Distribution:

```
stdNBdist(dataset["Addiction_Level"])
```

5.10702591327572e-17



Based on this image, most teens have addiction levels close to near average, few teens has extreme level.

Data Science-Bivariant Analysis

Covariance:

It tells about the difference between two columns. For example difference between Depression_Level & Sleep_Hours is **-0.049306.**This is **Negative Covariance** (ie) Depression_Level is increase Sleep_Hours is decrease and vice versa.

dataset[quan].cov()

	ID	Age	Daily_Usage_Hours	Sleep_Hours	Academic_Performance	Social_Interactions	Exercise_Hours	Anxiety_Level	Depression_
ID	750250.000000	-61.474658	-7.965622	23.569740	-172.841947	-111.841447	0.735278	22.057686	11.88
Age	-61.474658	3.958066	0.158380	0.043671	0.685964	-0.126413	0.001734	0.084258	0.30
Daily_Usage_Hours	-7.965622	0.158380	3.827896	0.047961	0.613226	-0.186614	-0.008523	-0.040840	0.05
Sleep_Hours	23.569740	0.043671	0.047961	2.222226	-0.005974	-0.083762	0.007462	0.042385	-0.04
cademic_Performance	-172.841947	0.685964	0.613226	-0.005974	215.624434	0.544659	-0.027067	0.145455	-1.11
Social_Interactions	-111.841447	-0.126413	-0.186614	-0.083762	0.544659	9.855413	0.013733	0.039390	0.21
Exercise_Hours	0.735278	0.001734	-0.008523	0.007462	-0.027067	0.013733	0.539666	0.009610	-0.02
Anxiety_Level	22.057686	0.084258	-0.040840	0.042385	0.145455	0.039390	0.009610	8.356019	0.15
Depression_Level	11.880460	0.301064	0.054338	-0.049306	-1.117795	0.218113	-0.024828	0.154455	8.24
Self_Esteem	-14.909470	-0.188824	0.035488	0.070314	-0.229303	0.074666	-0.030094	0.032674	-0.21
Parental_Control	-10.489496	0.028065	0.000949	0.004226	0.023394	-0.036562	0.007604	-0.016666	0.02
een_Time_Before_Bed	1.835812	0.007073	0.004759	-0.003169	-0.042159	-0.029501	0.006195	-0.006742	-0.04
hone_Checks_Per_Day	-1213.544348	-0.653063	0.350728	0.292516	-9.479595	1.421057	-0.361536	1.962784	-0.12
Apps_Used_Daily	-18.406802	-0.027526	0.206076	0.181729	-1.769832	-0.368301	-0.021220	0.098860	0.25
Time_on_Social_Media	15.397583	-0.010027	-0.024399	-0.026940	0.527569	0.008611	-0.013777	-0.007817	0.00
Time_on_Gaming	11.715138	-0.016639	-0.018925	0.008061	-0.462023	-0.014406	-0.004796	0.039573	-0.02
Time_on_Education	-0.407669	0.010799	0.018095	-0.009429	0.155912	-0.016801	0.008098	0.046179	-0.05
amily_Communication	-47.809770	-0.021731	0.024808	-0.115468	-1.171848	0.038119	0.013578	0.091827	-0.10
Weekend_Usage_Hours	32.430360	0.011228	0.077067	-0.011022	0.421969	-0.135420	0.059016	0.039137	-0.03
Addiction_Level	-10.126125	0.100251	1.891935	-0.519915	0.289877	-0.053717	-0.024849	0.074470	0.03

Correlation:

It tells about the relation between two columns.The relation between Time_on_Gaming and Addiction_level is **0.273060.**This is **positive correlation** (ie) Time_on_Gaming is increase Addiction_level is also increase.

dataset[quan].corr()

[1]:

	ID	Age	Daily_Usage_Hours	Sleep_Hours	Academic_Performance	Social_Interactions	Exercise_Hours	Anxiety_Level	Depressio
ID	1.000000	-0.035674	-0.004700	0.018254	-0.013589	-0.041130	0.001156	0.008810	0
Age	-0.035674	1.000000	0.040689	0.014725	0.023481	-0.020240	0.001187	0.014651	0
Daily_Usage_Hours	-0.004700	0.040689	1.000000	0.016444	0.021345	-0.030383	-0.005930	-0.007221	0
Sleep_Hours	0.018254	0.014725	0.016444	1.000000	-0.000273	-0.017898	0.006814	0.009836	-0
Academic_Performance	-0.013589	0.023481	0.021345	-0.000273	1.000000	0.011815	-0.002509	0.003427	-0
Social_Interactions	-0.041130	-0.020240	-0.030383	-0.017898	0.011815	1.000000	0.005955	0.004341	0
Exercise_Hours	0.001156	0.001187	-0.005930	0.006814	-0.002509	0.005955	1.000000	0.004525	-0
Anxiety_Level	0.008810	0.014651	-0.007221	0.009836	0.003427	0.004341	0.004525	1.000000	0
Depression_Level	0.004777	0.052699	0.009672	-0.011518	-0.026509	0.024195	-0.011770	0.018607	1
Self_Esteem	-0.006017	-0.033177	0.006340	0.016488	-0.005459	0.008314	-0.014320	0.003951	-0
Parental_Control	-0.024219	0.028212	0.000970	0.005670	0.003186	-0.023291	0.020701	-0.011530	0
Screen_Time_Before_Bed	0.004300	0.007213	0.004935	-0.004313	-0.005825	-0.019066	0.017109	-0.004732	-0
Phone_Checks_Per_Day	-0.037117	-0.008696	0.004749	0.005198	-0.017102	0.011992	-0.013038	0.017988	-0
Apps_Used_Daily	-0.004608	-0.003000	0.022841	0.026436	-0.026136	-0.025440	-0.006264	0.007416	0
Time_on_Social_Media	0.017989	-0.005100	-0.012620	-0.018288	0.036357	0.002776	-0.018977	-0.002736	0
Time_on_Gaming	0.014501	-0.008967	-0.010371	0.005798	-0.033734	-0.004920	-0.006999	0.014677	-0
Time_on_Education	-0.000726	0.008372	0.014265	-0.009756	0.016377	-0.008254	0.017003	0.024640	-0
Family_Communication	-0.019269	-0.003813	0.004427	-0.027040	-0.027859	0.004239	0.006452	0.011090	-0
Weekend_Usage_Hours	0.018583	0.002801	0.019551	-0.003670	0.014263	-0.021410	0.039873	0.006720	-0
Addiction_Level	-0.007263	0.031306	0.600771	-0.216681	0.012264	-0.010631	-0.021015	0.016005	0

Variance Inflation Factor(VIF):

To overcome Multicolinearity(two or more independent variable in linear model) by using **VIF.**

If VIF is

1 Not Correlated

Between 1&5 Moderately Correlated

Greater than 5 highly correlated

```
cal_vif(dataset[['Daily_Usage_Hours','Sleep_Hours','Exercise_Hours', 'Depression_Level', 'Anxiety_Level','Screen_Time_Before_Bed']])
```

	variables	VIF
0	Daily_Usage_Hours	6.410469
1	Sleep_Hours	10.620278
2	Exercise_Hours	2.864827
3	Depression_Level	4.174442
4	Anxiety_Level	4.323734
5	Screen_Time_Before_Bed	4.611200

Based on the output

Daily_usage_Hours is **highly correlated**

Sleep_Hours is **very highly correlated**

Exercise_hours,Depression_Level,Anxiety_Level,

Screen_Time_Before_Bed is **moderately correlated**

Ttest:

To find similarity between two groups based on **Mean.** It has two types

- Paired
- UnPaired

```
]: #DEPENDENTSAMPLE-PAIR
   #SAME GROUP BUT DIFFERENT CONDITION
```

```
]: male=dataset[dataset['Gender']=='Male']['Time_on_Social_Media']
   male1=dataset[dataset['Gender']=='Male']['Time_on_Education']
   ttest_rel(male,male1)
```

```
]: TtestResult(statistic=39.035728254652184, pvalue=2.7717864719557496e-204, df=1015)
```

Since the p-value < 0.05, you reject the null hypothesis.

Interpretation:

There is a statistically significant difference in the time that male spends on social media compared to the time they spend on education.

```
#INDEPENDENTSAMPLE-UNPAIR
#DIFFERENT GROUP BUT SAME CONDITION
```

```
grade9=dataset[dataset['School_Grade']=='9th']['Addiction_Level']
grade10=dataset[dataset['School_Grade']=='10th']['Addiction_Level']
ttest_ind(grade9,grade10)
```

```
TtestResult(statistic=0.2490729336822223, pvalue=0.8033549559450139, df=1011.0)
```

The p-value (0.803) is greater than 0.05, we fail to reject the null hypothesis. This means there is no statistically significant difference between the Addiction Levels of 9th and 10th grade students.

Analysis of Variance (ANAVO):

ANAVO method is used to compare the means of 3 or more groups to see atleast one group's mean is significantly different from the others.It has two types

- **One-way ANOVA**
- **Two-way ANOVA**

One-way:

```
stats.f_oneway(dataset['Daily_Usage_Hours'],dataset['Sleep_Hours'],dataset['Weekend_Usage_Hours'])
```

```
F_onewayResult(statistic=500.39267350735423, pvalue=8.736036274018552e-207)
```

```
#Reject Null hypothesis p<0.05
#Accept  Alternative hypothsis
#H0:There is no significance differance between these columns-Null Hypothesis
#H1:There is a significance differance between these columns-Alternative hypothesis
```

```
#pvalue is greater than 0.05 so we accept alternative hypothsis.Moreover there is a significance differance between these columns.
```

Two-way:

```
two_Way=ols('Addiction_Level ~ Gender + School_Grade + Location + Gender*  School_Grade',data=dataset).fit()
```

```
Two=sm.stats.anova_lm(two_Way,typ=2)
```

```
Two
```

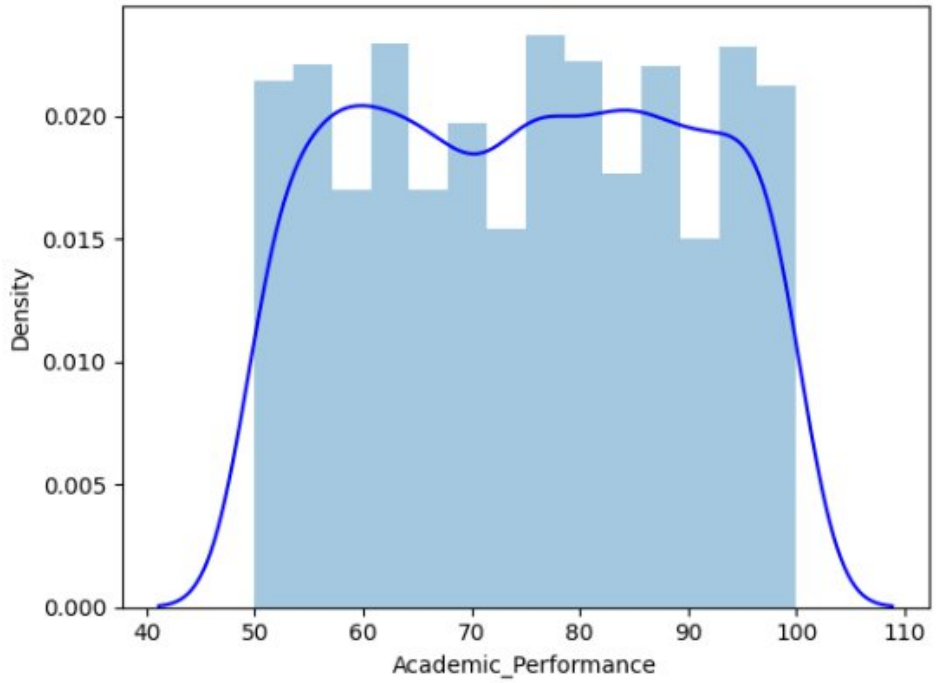
	sum_sq	df	F	PR(>F)
Gender	0.498121	2.0	0.105722	0.899714
School_Grade	17.996865	5.0	1.527875	0.181469
Location	7131.664300	2725.0	1.110927	0.136167
Gender:School_Grade	29.599487	10.0	1.256450	0.255668
Residual	605.441340	257.0	NaN	NaN

Data Visualization-Univariant Analysis

It represent the data in visual form and easy to understand the relationship between data.

Distribution plot:

```
]: sb.distplot(dataset['Academic_Performance'],hist=True,kde=True,kde_kws={'color':'blue'})
]: <Axes: xlabel='Academic_Performance', ylabel='Density'>
```

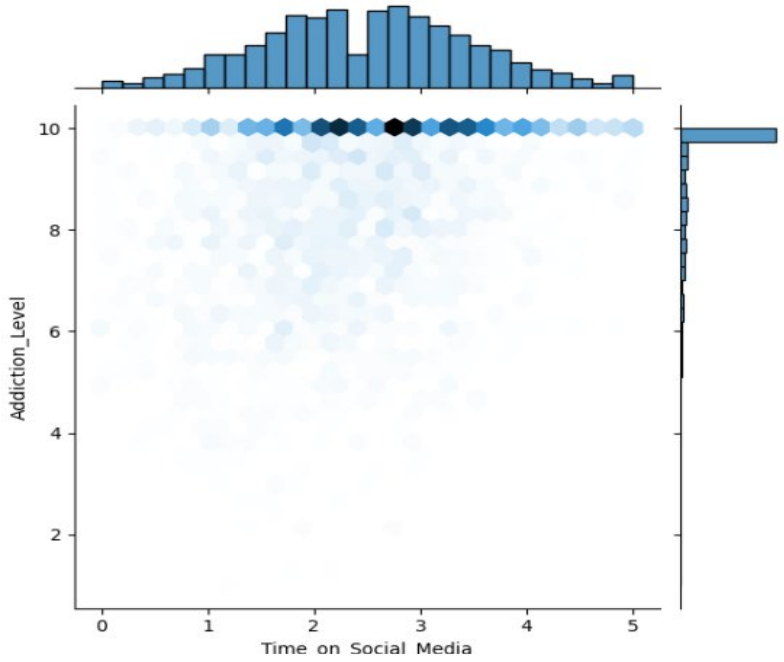


The graph visually demonstrates how "Academic Performance" is distributed across the dataset. Moreover peaks around 60-65, 75-80, and 90-95 that mean most of the students fall between these ranges.

Data Visualization-Bivariant Analysis

Joint plot:

```
sb.jointplot(x="Time_on_Social_Media",y="Addiction_Level",data=dataset,kind="hex")
plt.show()
```



In the above graph **Dark blue** indicate **more** students spending that specific amount of time on social media at that particular Addiction_Level. **Light blue** indicate **less** students spending that specific amount of time on social media at that particular Addiction_Level.

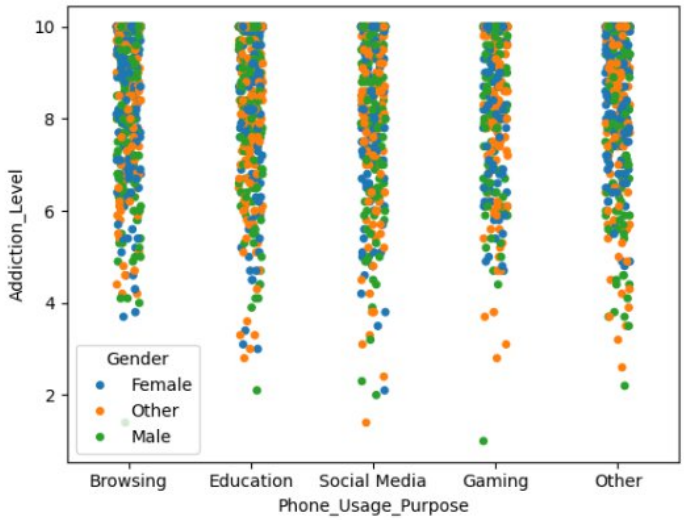
Top Side Histogram shows the distribution of Time on Social Media. Most of the students are grouped in the lower to mid range that is low to medium amount of time used in social media.

Right Side Histogram shows the distribution of Addiction_Level.Many students have higher addiction level (close to 10).

Strip plot:(for Categorial Data)

```
#For CATEGORICAL DATA

sb.stripplot(x="Phone_Usage_Purpose",y="Addiction_Level",data=dataset,hue="Gender")
plt.show()
```



In the above graph **Browsing, Social Media, Gaming** --> most Addictional_Level are pointed between 5 to 10.**Education** is less extreme level compared to Browsing, Social Media, Gaming.

Other purposes the pointes are more widely spread out, with only a few cases reaching the maximum Addiction Level of 10.Both male and female teenagers are approximately same Addiction Level.

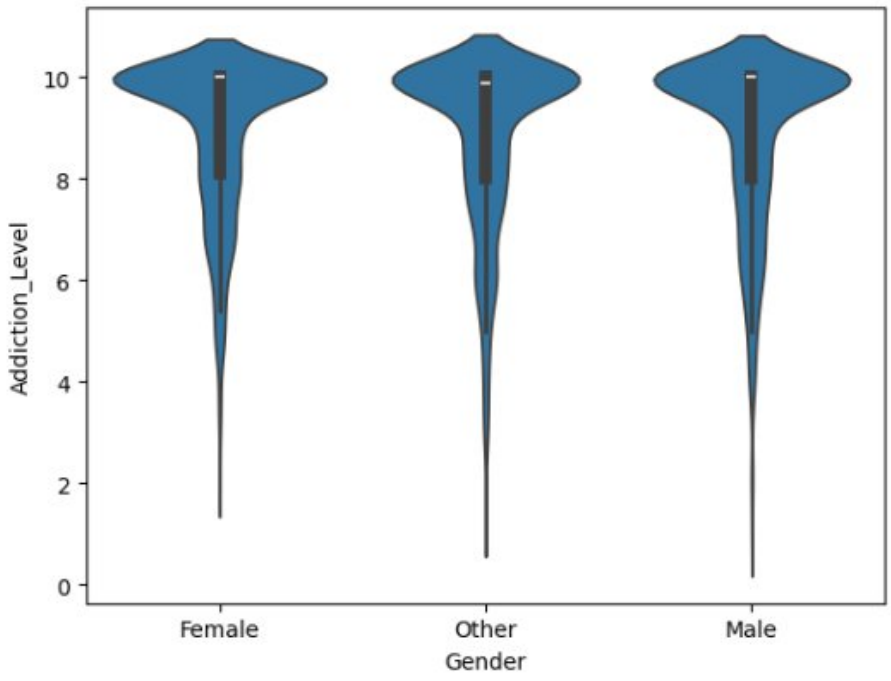
Violin plot:(for Categorial Data vs Numercial Data)

In the following graph **Wider part of blue area** indicate more individuals have that addiction level. **Narrow part** indicate fewer individuals at that level.**White dot** inside the violin shape indicate the average value. **Thick Black bar** indicate IQR (middle 50% of the data).

All the gender(male,female,other)have similar distribution level.**Most addiction levels are concentrated between 6 and 9**,showing high tendency toward phone addiction.Very few cases are near 0-2 or 10

```
15]: #Distribution plot-categoricaldata vs numercialdata

16]: sb.violinplot(x="Gender",y="Addiction_Level",data=dataset)
     plt.show()
```



Data Studio-Tableau Public

1. What is the **Maximum** daily phone usage time for each gender categorized by grade?
2. How does the **minimum** addiction level vary by grade?
3. What is the **median** depression level based on gender?
4. What is the **25th percentile** of addiction level based on phone usage purpose?
5. What is the **total count** of exercise hours and phone checks per day?
6. Which gender spends the **most time** on social media and education?