**#feature extraction, pre-processing and predicting the absenteeism of employees at work**

**import** pandas **as** pd

**import** matplotlib.pyplot **as** plt

**import** numpy **as** np

df**=**pd.read\_csv('C:/Users/Nachiketa/Desktop/ml1/absent\_uciv\_ass2/Absenteeism\_at\_work.csv', delimiter**=**";")

df

df.head(10)

​

Out[2]:

|  | **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** | **Social smoker** | **Pet** | **Weight** | **Height** | **Body mass index** | **Absenteeism time in hours** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 11 | 26 | 7 | 3 | 1 | 289 | 36 | 13 | 33 | 239.554 | ... | 0 | 1 | 2 | 1 | 0 | 1 | 90 | 172 | 30 | 4 |
| **1** | 36 | 0 | 7 | 3 | 1 | 118 | 13 | 18 | 50 | 239.554 | ... | 1 | 1 | 1 | 1 | 0 | 0 | 98 | 178 | 31 | 0 |
| **2** | 3 | 23 | 7 | 4 | 1 | 179 | 51 | 18 | 38 | 239.554 | ... | 0 | 1 | 0 | 1 | 0 | 0 | 89 | 170 | 31 | 2 |
| **3** | 7 | 7 | 7 | 5 | 1 | 279 | 5 | 14 | 39 | 239.554 | ... | 0 | 1 | 2 | 1 | 1 | 0 | 68 | 168 | 24 | 4 |
| **4** | 11 | 23 | 7 | 5 | 1 | 289 | 36 | 13 | 33 | 239.554 | ... | 0 | 1 | 2 | 1 | 0 | 1 | 90 | 172 | 30 | 2 |
| **5** | 3 | 23 | 7 | 6 | 1 | 179 | 51 | 18 | 38 | 239.554 | ... | 0 | 1 | 0 | 1 | 0 | 0 | 89 | 170 | 31 | 2 |
| **6** | 10 | 22 | 7 | 6 | 1 | 361 | 52 | 3 | 28 | 239.554 | ... | 0 | 1 | 1 | 1 | 0 | 4 | 80 | 172 | 27 | 8 |
| **7** | 20 | 23 | 7 | 6 | 1 | 260 | 50 | 11 | 36 | 239.554 | ... | 0 | 1 | 4 | 1 | 0 | 0 | 65 | 168 | 23 | 4 |
| **8** | 14 | 19 | 7 | 2 | 1 | 155 | 12 | 14 | 34 | 239.554 | ... | 0 | 1 | 2 | 1 | 0 | 0 | 95 | 196 | 25 | 40 |
| **9** | 1 | 22 | 7 | 2 | 1 | 235 | 11 | 14 | 37 | 239.554 | ... | 0 | 3 | 1 | 0 | 0 | 1 | 88 | 172 | 29 | 8 |

10 rows × 21 columns

In [3]:



print('Shape of dataset is:{}'.format(df.shape))

print('Type of features is:{}'.format(df.dtypes))

Shape of dataset is:(740, 21)

Type of features is:ID int64

Reason for absence int64

Month of absence int64

Day of the week int64

Seasons int64

Transportation expense int64

Distance from Residence to Work int64

Service time int64

Age int64

Work load Average/day float64

Hit target int64

Disciplinary failure int64

Education int64

Son int64

Social drinker int64

Social smoker int64

Pet int64

Weight int64

Height int64

Body mass index int64

Absenteeism time in hours int64

dtype: object

In [4]:



df['Disciplinary failure'].mean()

​

Out[4]:

0.05405405405405406

In [5]:



df['Absenteeism time in hours'].mean()

Out[5]:

6.924324324324324

In [6]:



**import** seaborn **as** sns

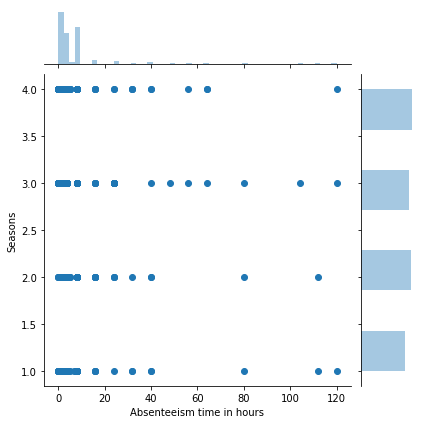
sns.jointplot(x**=**'Absenteeism time in hours',y**=**'Seasons',data**=**df)

C:\Users\Nachiketa\Anaconda3\lib\site-packages\scipy\stats\stats.py:1713: FutureWarning: Using a non-tuple sequence for multidimensional indexing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interpreted as an array index, `arr[np.array(seq)]`, which will result either in an error or a different result.

return np.add.reduce(sorted[indexer] \* weights, axis=axis) / sumval

Out[6]:

<seaborn.axisgrid.JointGrid at 0x176f3b9c780>



In [23]:



**import** numpy **as** np

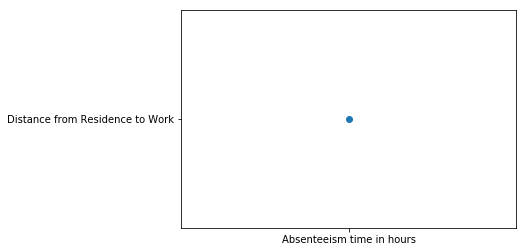
x**=**'Absenteeism time in hours'

y**=**'Distance from Residence to Work'

plt.scatter(x,y)

Out[23]:

<matplotlib.collections.PathCollection at 0x22f747307b8>



In [7]:



plt.scatter(x**=**'Absenteeism time in hours',y**=**'Disciplinary failure')

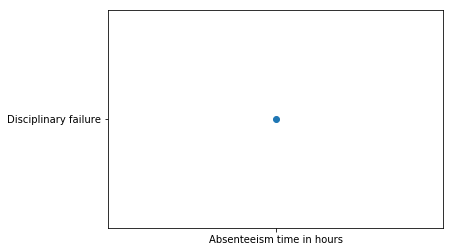
sns.jointplot(x**=**'Absenteeism time in hours',y**=**'Disciplinary failure',data**=**df)

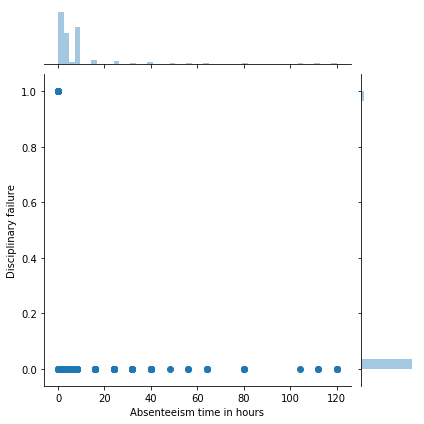
C:\Users\Nachiketa\Anaconda3\lib\site-packages\scipy\stats\stats.py:1713: FutureWarning: Using a non-tuple sequence for multidimensional indexing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interpreted as an array index, `arr[np.array(seq)]`, which will result either in an error or a different result.

return np.add.reduce(sorted[indexer] \* weights, axis=axis) / sumval

Out[7]:

<seaborn.axisgrid.JointGrid at 0x176f4c2c940>





In [8]:



df.tail(10)

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** | **Social smoker** | **Pet** | **Weight** | **Height** | **Body mass index** | **Absenteeism time in hours** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **730** | 6 | 22 | 7 | 3 | 1 | 189 | 29 | 13 | 33 | 264.604 | ... | 0 | 1 | 2 | 0 | 0 | 2 | 69 | 167 | 25 | 16 |
| **731** | 34 | 23 | 7 | 4 | 1 | 118 | 10 | 10 | 37 | 264.604 | ... | 0 | 1 | 0 | 0 | 0 | 0 | 83 | 172 | 28 | 2 |
| **732** | 10 | 22 | 7 | 4 | 1 | 361 | 52 | 3 | 28 | 264.604 | ... | 0 | 1 | 1 | 1 | 0 | 4 | 80 | 172 | 27 | 8 |
| **733** | 28 | 22 | 7 | 4 | 1 | 225 | 26 | 9 | 28 | 264.604 | ... | 0 | 1 | 1 | 0 | 0 | 2 | 69 | 169 | 24 | 8 |
| **734** | 13 | 13 | 7 | 2 | 1 | 369 | 17 | 12 | 31 | 264.604 | ... | 0 | 1 | 3 | 1 | 0 | 0 | 70 | 169 | 25 | 80 |
| **735** | 11 | 14 | 7 | 3 | 1 | 289 | 36 | 13 | 33 | 264.604 | ... | 0 | 1 | 2 | 1 | 0 | 1 | 90 | 172 | 30 | 8 |
| **736** | 1 | 11 | 7 | 3 | 1 | 235 | 11 | 14 | 37 | 264.604 | ... | 0 | 3 | 1 | 0 | 0 | 1 | 88 | 172 | 29 | 4 |
| **737** | 4 | 0 | 0 | 3 | 1 | 118 | 14 | 13 | 40 | 271.219 | ... | 0 | 1 | 1 | 1 | 0 | 8 | 98 | 170 | 34 | 0 |
| **738** | 8 | 0 | 0 | 4 | 2 | 231 | 35 | 14 | 39 | 271.219 | ... | 0 | 1 | 2 | 1 | 0 | 2 | 100 | 170 | 35 | 0 |
| **739** | 35 | 0 | 0 | 6 | 3 | 179 | 45 | 14 | 53 | 271.219 | ... | 0 | 1 | 1 | 0 | 0 | 1 | 77 | 175 | 25 | 0 |

10 rows × 21 columns

In [9]:



plt.figure(figsize**=**(20,10))

sns.lmplot(x**=**'Age',y**=**'Absenteeism time in hours',data**=**df,hue**=**'Day of the week',size**=**20,aspect**=**1)

​

C:\Users\Nachiketa\Anaconda3\lib\site-packages\seaborn\regression.py:546: UserWarning: The `size` paramter has been renamed to `height`; please update your code.

warnings.warn(msg, UserWarning)

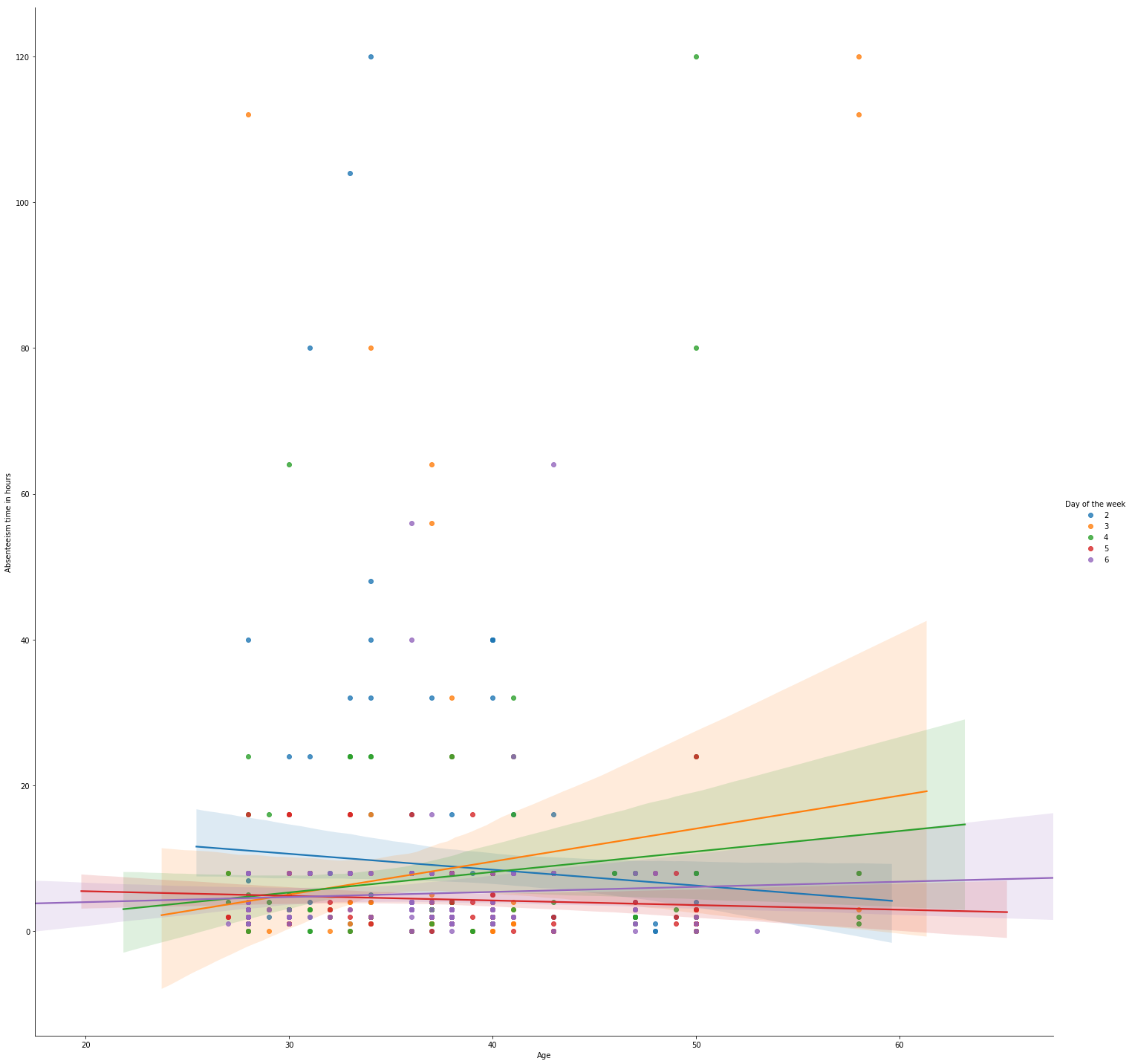
C:\Users\Nachiketa\Anaconda3\lib\site-packages\scipy\stats\stats.py:1713: FutureWarning: Using a non-tuple sequence for multidimensional indexing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interpreted as an array index, `arr[np.array(seq)]`, which will result either in an error or a different result.

return np.add.reduce(sorted[indexer] \* weights, axis=axis) / sumval

Out[9]:

<seaborn.axisgrid.FacetGrid at 0x176f5e502e8>

<Figure size 1440x720 with 0 Axes>



In [10]:



​

df[df['Day of the week']**==**3]['Absenteeism time in hours'].mean()

Out[10]:

7.98051948051948

In [11]:



df[df['Month of absence']**==**7]['Absenteeism time in hours'].mean()

Out[11]:

10.955223880597014

In [12]:



plt.figure(figsize**=**(20,10))

sns.lmplot(x**=**'Month of absence',y**=**'Absenteeism time in hours',data**=**df,hue**=**'Day of the week',size**=**20,aspect**=**1)

​

C:\Users\Nachiketa\Anaconda3\lib\site-packages\seaborn\regression.py:546: UserWarning: The `size` paramter has been renamed to `height`; please update your code.

warnings.warn(msg, UserWarning)

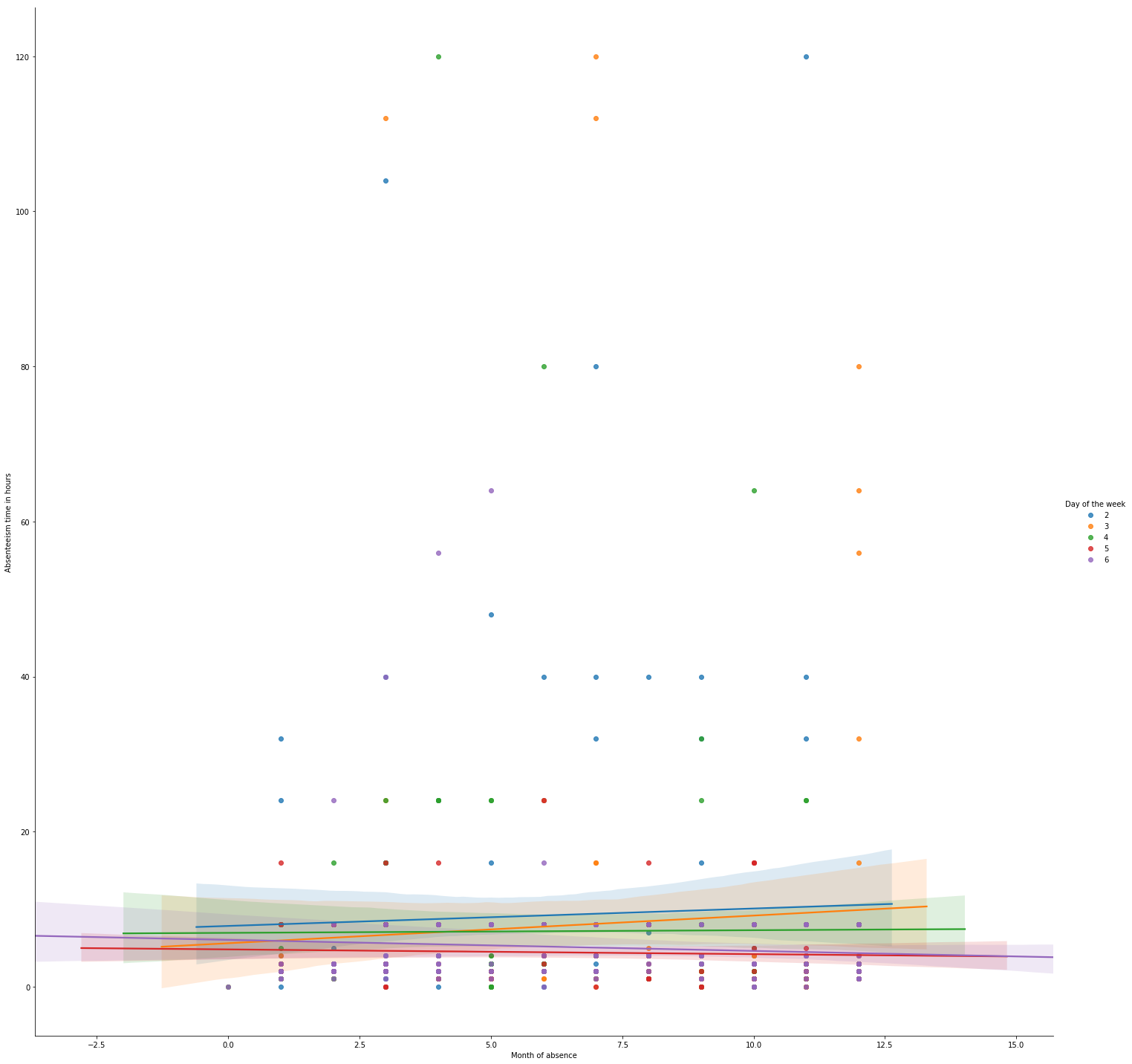
C:\Users\Nachiketa\Anaconda3\lib\site-packages\scipy\stats\stats.py:1713: FutureWarning: Using a non-tuple sequence for multidimensional indexing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interpreted as an array index, `arr[np.array(seq)]`, which will result either in an error or a different result.

return np.add.reduce(sorted[indexer] \* weights, axis=axis) / sumval

Out[12]:

<seaborn.axisgrid.FacetGrid at 0x176f674f390>

<Figure size 1440x720 with 0 Axes>



In [13]:



df['Transportation expense'].mean()

Out[13]:

221.32972972972973

In [15]:



sns.jointplot(x**=**'Absenteeism time in hours',y**=**'Transportation expense',data**=**df,color**=**'yellow')

​

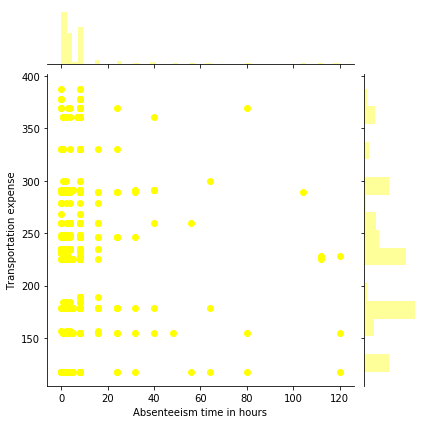
sns.jointplot(x**=**'Transportation expense',y**=**'Month of absence',data**=**df,kind**=**'hex',color**=**'brown')

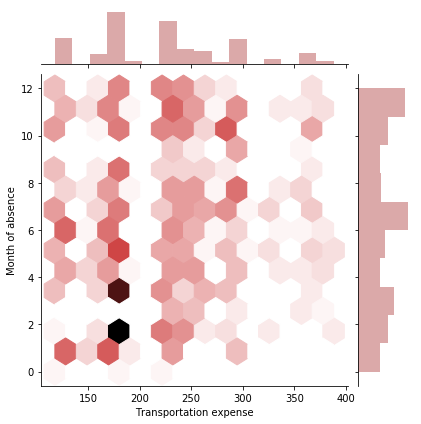
C:\Users\Nachiketa\Anaconda3\lib\site-packages\scipy\stats\stats.py:1713: FutureWarning: Using a non-tuple sequence for multidimensional indexing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interpreted as an array index, `arr[np.array(seq)]`, which will result either in an error or a different result.

return np.add.reduce(sorted[indexer] \* weights, axis=axis) / sumval

Out[15]:

<seaborn.axisgrid.JointGrid at 0x176f721e630>





In [16]:



df.head(10)

df['Son'].value\_counts()

​

Out[16]:

0 298

1 229

2 156

4 42

3 15

Name: Son, dtype: int64

In [17]:



df.head()

Out[17]:

|  | **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** | **Social smoker** | **Pet** | **Weight** | **Height** | **Body mass index** | **Absenteeism time in hours** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 11 | 26 | 7 | 3 | 1 | 289 | 36 | 13 | 33 | 239.554 | ... | 0 | 1 | 2 | 1 | 0 | 1 | 90 | 172 | 30 | 4 |
| **1** | 36 | 0 | 7 | 3 | 1 | 118 | 13 | 18 | 50 | 239.554 | ... | 1 | 1 | 1 | 1 | 0 | 0 | 98 | 178 | 31 | 0 |
| **2** | 3 | 23 | 7 | 4 | 1 | 179 | 51 | 18 | 38 | 239.554 | ... | 0 | 1 | 0 | 1 | 0 | 0 | 89 | 170 | 31 | 2 |
| **3** | 7 | 7 | 7 | 5 | 1 | 279 | 5 | 14 | 39 | 239.554 | ... | 0 | 1 | 2 | 1 | 1 | 0 | 68 | 168 | 24 | 4 |
| **4** | 11 | 23 | 7 | 5 | 1 | 289 | 36 | 13 | 33 | 239.554 | ... | 0 | 1 | 2 | 1 | 0 | 1 | 90 | 172 | 30 | 2 |

5 rows × 21 columns

In [18]:



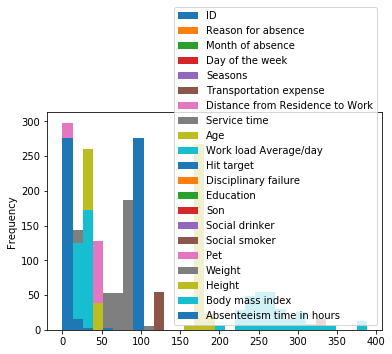
plt.figure(figsize**=**(20,10))

df[df['Son']**==**0].plot.hist(bins**=**30)

Out[18]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x176f7d0f940>

<Figure size 1440x720 with 0 Axes>



In [19]:

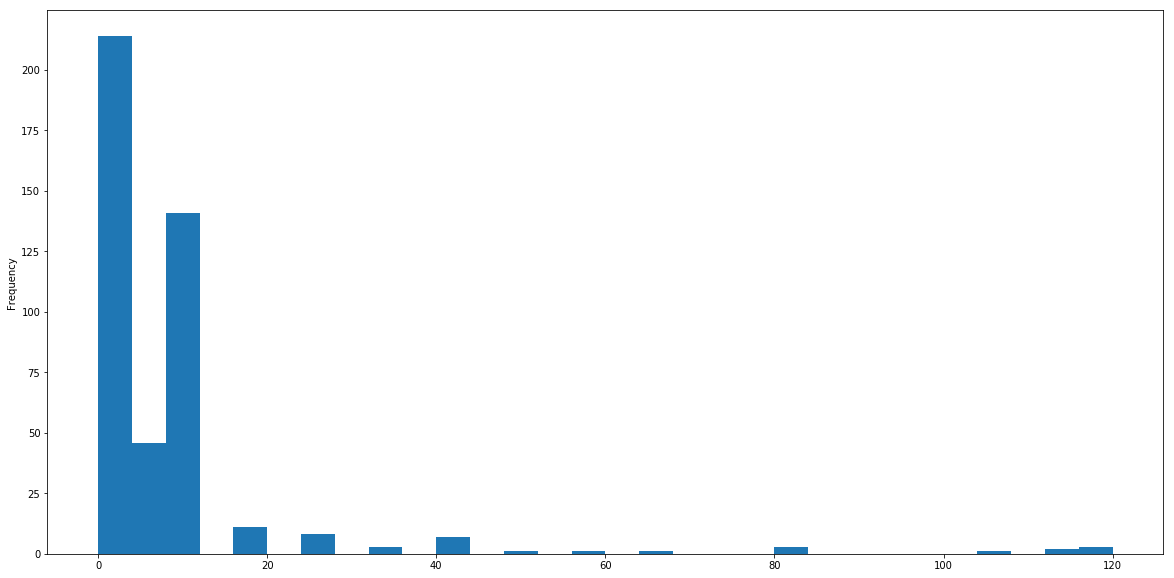


plt.figure(figsize**=**(20,10))

df[df['Son']**!=**0]['Absenteeism time in hours'].plot.hist(bins**=**30)

Out[19]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x176f7edd5c0>



In [20]:

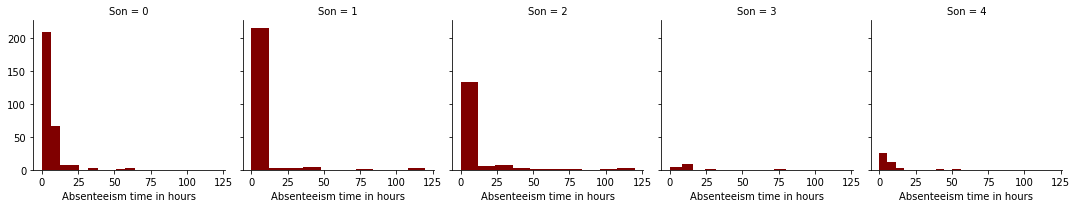


g **=** sns.FacetGrid(data**=**df,col**=**'Son')

g.map(plt.hist,'Absenteeism time in hours',color**=**'maroon')

Out[20]:

<seaborn.axisgrid.FacetGrid at 0x176f801bc50>



In [21]:



g **=** sns.FacetGrid(data**=**df,col**=**'Service time',size**=**5)

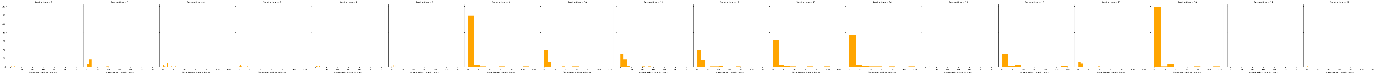
g.map(plt.hist,'Absenteeism time in hours',color**=**'orange')

C:\Users\Nachiketa\Anaconda3\lib\site-packages\seaborn\axisgrid.py:230: UserWarning: The `size` paramter has been renamed to `height`; please update your code.

warnings.warn(msg, UserWarning)

Out[21]:

<seaborn.axisgrid.FacetGrid at 0x176fa2b4d68>



In [22]:



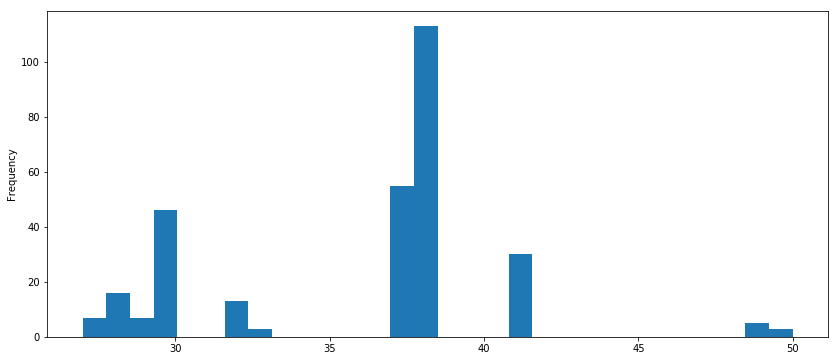
plt.figure(figsize**=**(14,6))

df[df['Son']**==**0]['Age'].plot.hist(bins**=**30)

​

Out[22]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x176f8450d68>



In [23]:



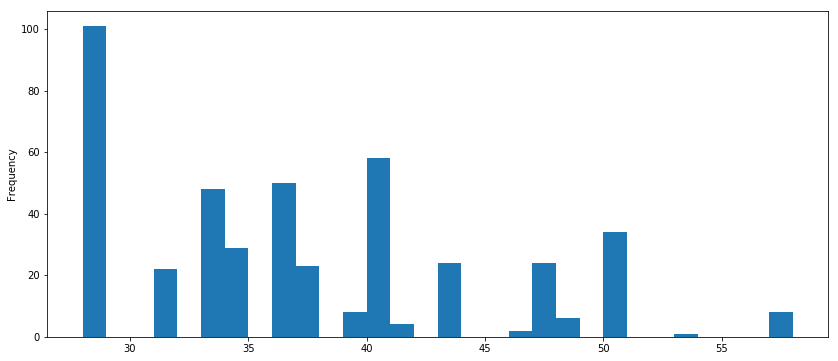
plt.figure(figsize**=**(14,6))

df[df['Son']**!=**0]['Age'].plot.hist(bins**=**30)

​

Out[23]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x176fb90b278>



In [24]:



reasons **=** df['Reason for absence']

plt.figure(figsize**=**(10,5))

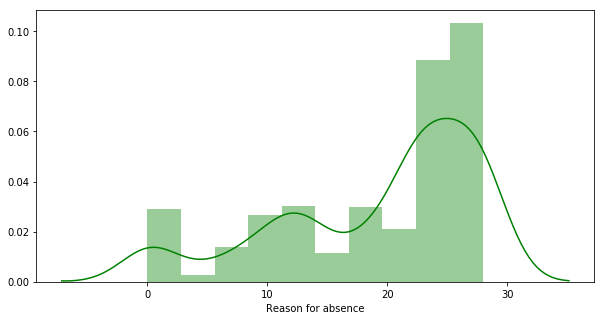
sns.distplot(df['Reason for absence'],color**=**'green')

C:\Users\Nachiketa\Anaconda3\lib\site-packages\scipy\stats\stats.py:1713: FutureWarning: Using a non-tuple sequence for multidimensional indexing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interpreted as an array index, `arr[np.array(seq)]`, which will result either in an error or a different result.

return np.add.reduce(sorted[indexer] \* weights, axis=axis) / sumval

Out[24]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x176fb966cc0>



In [25]:



df.tail()

df.head()

df.head(700)

Out[25]:

|  | **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** | **Social smoker** | **Pet** | **Weight** | **Height** | **Body mass index** | **Absenteeism time in hours** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 11 | 26 | 7 | 3 | 1 | 289 | 36 | 13 | 33 | 239.554 | ... | 0 | 1 | 2 | 1 | 0 | 1 | 90 | 172 | 30 | 4 |
| **1** | 36 | 0 | 7 | 3 | 1 | 118 | 13 | 18 | 50 | 239.554 | ... | 1 | 1 | 1 | 1 | 0 | 0 | 98 | 178 | 31 | 0 |
| **2** | 3 | 23 | 7 | 4 | 1 | 179 | 51 | 18 | 38 | 239.554 | ... | 0 | 1 | 0 | 1 | 0 | 0 | 89 | 170 | 31 | 2 |
| **3** | 7 | 7 | 7 | 5 | 1 | 279 | 5 | 14 | 39 | 239.554 | ... | 0 | 1 | 2 | 1 | 1 | 0 | 68 | 168 | 24 | 4 |
| **4** | 11 | 23 | 7 | 5 | 1 | 289 | 36 | 13 | 33 | 239.554 | ... | 0 | 1 | 2 | 1 | 0 | 1 | 90 | 172 | 30 | 2 |
| **5** | 3 | 23 | 7 | 6 | 1 | 179 | 51 | 18 | 38 | 239.554 | ... | 0 | 1 | 0 | 1 | 0 | 0 | 89 | 170 | 31 | 2 |
| **6** | 10 | 22 | 7 | 6 | 1 | 361 | 52 | 3 | 28 | 239.554 | ... | 0 | 1 | 1 | 1 | 0 | 4 | 80 | 172 | 27 | 8 |
| **7** | 20 | 23 | 7 | 6 | 1 | 260 | 50 | 11 | 36 | 239.554 | ... | 0 | 1 | 4 | 1 | 0 | 0 | 65 | 168 | 23 | 4 |
| **8** | 14 | 19 | 7 | 2 | 1 | 155 | 12 | 14 | 34 | 239.554 | ... | 0 | 1 | 2 | 1 | 0 | 0 | 95 | 196 | 25 | 40 |
| **9** | 1 | 22 | 7 | 2 | 1 | 235 | 11 | 14 | 37 | 239.554 | ... | 0 | 3 | 1 | 0 | 0 | 1 | 88 | 172 | 29 | 8 |
| **10** | 20 | 1 | 7 | 2 | 1 | 260 | 50 | 11 | 36 | 239.554 | ... | 0 | 1 | 4 | 1 | 0 | 0 | 65 | 168 | 23 | 8 |
| **11** | 20 | 1 | 7 | 3 | 1 | 260 | 50 | 11 | 36 | 239.554 | ... | 0 | 1 | 4 | 1 | 0 | 0 | 65 | 168 | 23 | 8 |
| **12** | 20 | 11 | 7 | 4 | 1 | 260 | 50 | 11 | 36 | 239.554 | ... | 0 | 1 | 4 | 1 | 0 | 0 | 65 | 168 | 23 | 8 |
| **13** | 3 | 11 | 7 | 4 | 1 | 179 | 51 | 18 | 38 | 239.554 | ... | 0 | 1 | 0 | 1 | 0 | 0 | 89 | 170 | 31 | 1 |
| **14** | 3 | 23 | 7 | 4 | 1 | 179 | 51 | 18 | 38 | 239.554 | ... | 0 | 1 | 0 | 1 | 0 | 0 | 89 | 170 | 31 | 4 |
| **15** | 24 | 14 | 7 | 6 | 1 | 246 | 25 | 16 | 41 | 239.554 | ... | 0 | 1 | 0 | 1 | 0 | 0 | 67 | 170 | 23 | 8 |
| **16** | 3 | 23 | 7 | 6 | 1 | 179 | 51 | 18 | 38 | 239.554 | ... | 0 | 1 | 0 | 1 | 0 | 0 | 89 | 170 | 31 | 2 |
| **17** | 3 | 21 | 7 | 2 | 1 | 179 | 51 | 18 | 38 | 239.554 | ... | 0 | 1 | 0 | 1 | 0 | 0 | 89 | 170 | 31 | 8 |
| **18** | 6 | 11 | 7 | 5 | 1 | 189 | 29 | 13 | 33 | 239.554 | ... | 0 | 1 | 2 | 0 | 0 | 2 | 69 | 167 | 25 | 8 |
| **19** | 33 | 23 | 8 | 4 | 1 | 248 | 25 | 14 | 47 | 205.917 | ... | 0 | 1 | 2 | 0 | 0 | 1 | 86 | 165 | 32 | 2 |
| **20** | 18 | 10 | 8 | 4 | 1 | 330 | 16 | 4 | 28 | 205.917 | ... | 0 | 2 | 0 | 0 | 0 | 0 | 84 | 182 | 25 | 8 |
| **21** | 3 | 11 | 8 | 2 | 1 | 179 | 51 | 18 | 38 | 205.917 | ... | 0 | 1 | 0 | 1 | 0 | 0 | 89 | 170 | 31 | 1 |
| **22** | 10 | 13 | 8 | 2 | 1 | 361 | 52 | 3 | 28 | 205.917 | ... | 0 | 1 | 1 | 1 | 0 | 4 | 80 | 172 | 27 | 40 |
| **23** | 20 | 28 | 8 | 6 | 1 | 260 | 50 | 11 | 36 | 205.917 | ... | 0 | 1 | 4 | 1 | 0 | 0 | 65 | 168 | 23 | 4 |
| **24** | 11 | 18 | 8 | 2 | 1 | 289 | 36 | 13 | 33 | 205.917 | ... | 0 | 1 | 2 | 1 | 0 | 1 | 90 | 172 | 30 | 8 |
| **25** | 10 | 25 | 8 | 2 | 1 | 361 | 52 | 3 | 28 | 205.917 | ... | 0 | 1 | 1 | 1 | 0 | 4 | 80 | 172 | 27 | 7 |
| **26** | 11 | 23 | 8 | 3 | 1 | 289 | 36 | 13 | 33 | 205.917 | ... | 0 | 1 | 2 | 1 | 0 | 1 | 90 | 172 | 30 | 1 |
| **27** | 30 | 28 | 8 | 4 | 1 | 157 | 27 | 6 | 29 | 205.917 | ... | 0 | 1 | 0 | 1 | 1 | 0 | 75 | 185 | 22 | 4 |
| **28** | 11 | 18 | 8 | 4 | 1 | 289 | 36 | 13 | 33 | 205.917 | ... | 0 | 1 | 2 | 1 | 0 | 1 | 90 | 172 | 30 | 8 |
| **29** | 3 | 23 | 8 | 6 | 1 | 179 | 51 | 18 | 38 | 205.917 | ... | 0 | 1 | 0 | 1 | 0 | 0 | 89 | 170 | 31 | 2 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **670** | 14 | 28 | 4 | 3 | 3 | 155 | 12 | 14 | 34 | 246.288 | ... | 0 | 1 | 2 | 1 | 0 | 0 | 95 | 196 | 25 | 4 |
| **671** | 28 | 19 | 4 | 5 | 3 | 225 | 26 | 9 | 28 | 246.288 | ... | 0 | 1 | 1 | 0 | 0 | 2 | 69 | 169 | 24 | 8 |
| **672** | 36 | 14 | 4 | 5 | 3 | 118 | 13 | 18 | 50 | 246.288 | ... | 0 | 1 | 1 | 1 | 0 | 0 | 98 | 178 | 31 | 2 |
| **673** | 22 | 27 | 4 | 6 | 3 | 179 | 26 | 9 | 30 | 246.288 | ... | 0 | 3 | 0 | 0 | 0 | 0 | 56 | 171 | 19 | 2 |
| **674** | 1 | 22 | 5 | 2 | 3 | 235 | 11 | 14 | 37 | 237.656 | ... | 0 | 3 | 1 | 0 | 0 | 1 | 88 | 172 | 29 | 8 |
| **675** | 29 | 19 | 5 | 4 | 3 | 225 | 15 | 15 | 41 | 237.656 | ... | 0 | 4 | 2 | 1 | 0 | 2 | 94 | 182 | 28 | 3 |
| **676** | 25 | 28 | 5 | 4 | 3 | 235 | 16 | 8 | 32 | 237.656 | ... | 0 | 3 | 0 | 0 | 0 | 0 | 75 | 178 | 25 | 2 |
| **677** | 34 | 8 | 5 | 4 | 3 | 118 | 10 | 10 | 37 | 237.656 | ... | 0 | 1 | 0 | 0 | 0 | 0 | 83 | 172 | 28 | 3 |
| **678** | 5 | 26 | 5 | 4 | 3 | 235 | 20 | 13 | 43 | 237.656 | ... | 0 | 1 | 1 | 1 | 0 | 0 | 106 | 167 | 38 | 8 |
| **679** | 22 | 13 | 5 | 5 | 3 | 179 | 26 | 9 | 30 | 237.656 | ... | 0 | 3 | 0 | 0 | 0 | 0 | 56 | 171 | 19 | 1 |
| **680** | 15 | 28 | 5 | 5 | 3 | 291 | 31 | 12 | 40 | 237.656 | ... | 0 | 1 | 1 | 1 | 0 | 1 | 73 | 171 | 25 | 2 |
| **681** | 29 | 14 | 5 | 5 | 3 | 225 | 15 | 15 | 41 | 237.656 | ... | 0 | 4 | 2 | 1 | 0 | 2 | 94 | 182 | 28 | 8 |
| **682** | 26 | 19 | 5 | 6 | 3 | 300 | 26 | 13 | 43 | 237.656 | ... | 0 | 1 | 2 | 1 | 1 | 1 | 77 | 175 | 25 | 64 |
| **683** | 29 | 22 | 5 | 6 | 3 | 225 | 15 | 15 | 41 | 237.656 | ... | 0 | 4 | 2 | 1 | 0 | 2 | 94 | 182 | 28 | 8 |
| **684** | 22 | 27 | 5 | 6 | 3 | 179 | 26 | 9 | 30 | 237.656 | ... | 0 | 3 | 0 | 0 | 0 | 0 | 56 | 171 | 19 | 2 |
| **685** | 36 | 23 | 5 | 2 | 3 | 118 | 13 | 18 | 50 | 237.656 | ... | 0 | 1 | 1 | 1 | 0 | 0 | 98 | 178 | 31 | 2 |
| **686** | 36 | 5 | 5 | 3 | 3 | 118 | 13 | 18 | 50 | 237.656 | ... | 0 | 1 | 1 | 1 | 0 | 0 | 98 | 178 | 31 | 3 |
| **687** | 34 | 28 | 5 | 3 | 3 | 118 | 10 | 10 | 37 | 237.656 | ... | 0 | 1 | 0 | 0 | 0 | 0 | 83 | 172 | 28 | 1 |
| **688** | 36 | 0 | 5 | 3 | 3 | 118 | 13 | 18 | 50 | 237.656 | ... | 1 | 1 | 1 | 1 | 0 | 0 | 98 | 178 | 31 | 0 |
| **689** | 22 | 27 | 5 | 4 | 3 | 179 | 26 | 9 | 30 | 237.656 | ... | 0 | 3 | 0 | 0 | 0 | 0 | 56 | 171 | 19 | 2 |
| **690** | 23 | 0 | 5 | 4 | 3 | 378 | 49 | 11 | 36 | 237.656 | ... | 1 | 1 | 2 | 0 | 1 | 4 | 65 | 174 | 21 | 0 |
| **691** | 17 | 16 | 5 | 6 | 3 | 179 | 22 | 17 | 40 | 237.656 | ... | 0 | 2 | 2 | 0 | 1 | 0 | 63 | 170 | 22 | 1 |
| **692** | 14 | 10 | 5 | 2 | 3 | 155 | 12 | 14 | 34 | 237.656 | ... | 0 | 1 | 2 | 1 | 0 | 0 | 95 | 196 | 25 | 48 |
| **693** | 25 | 10 | 5 | 2 | 3 | 235 | 16 | 8 | 32 | 237.656 | ... | 0 | 3 | 0 | 0 | 0 | 0 | 75 | 178 | 25 | 8 |
| **694** | 15 | 22 | 5 | 4 | 3 | 291 | 31 | 12 | 40 | 237.656 | ... | 0 | 1 | 1 | 1 | 0 | 1 | 73 | 171 | 25 | 8 |
| **695** | 17 | 10 | 5 | 4 | 3 | 179 | 22 | 17 | 40 | 237.656 | ... | 0 | 2 | 2 | 0 | 1 | 0 | 63 | 170 | 22 | 8 |
| **696** | 28 | 6 | 5 | 4 | 3 | 225 | 26 | 9 | 28 | 237.656 | ... | 0 | 1 | 1 | 0 | 0 | 2 | 69 | 169 | 24 | 3 |
| **697** | 18 | 10 | 5 | 5 | 3 | 330 | 16 | 4 | 28 | 237.656 | ... | 0 | 2 | 0 | 0 | 0 | 0 | 84 | 182 | 25 | 8 |
| **698** | 25 | 23 | 5 | 5 | 3 | 235 | 16 | 8 | 32 | 237.656 | ... | 0 | 3 | 0 | 0 | 0 | 0 | 75 | 178 | 25 | 2 |
| **699** | 15 | 28 | 5 | 5 | 3 | 291 | 31 | 12 | 40 | 237.656 | ... | 0 | 1 | 1 | 1 | 0 | 1 | 73 | 171 | 25 | 2 |

700 rows × 21 columns

In [26]:



df[df['Reason for absence']**==**23].count()

Out[26]:

ID 149

Reason for absence 149

Month of absence 149

Day of the week 149

Seasons 149

Transportation expense 149

Distance from Residence to Work 149

Service time 149

Age 149

Work load Average/day 149

Hit target 149

Disciplinary failure 149

Education 149

Son 149

Social drinker 149

Social smoker 149

Pet 149

Weight 149

Height 149

Body mass index 149

Absenteeism time in hours 149

dtype: int64

In [27]:



df[df['Absenteeism time in hours']**==**1].count()

Out[27]:

ID 88

Reason for absence 88

Month of absence 88

Day of the week 88

Seasons 88

Transportation expense 88

Distance from Residence to Work 88

Service time 88

Age 88

Work load Average/day 88

Hit target 88

Disciplinary failure 88

Education 88

Son 88

Social drinker 88

Social smoker 88

Pet 88

Weight 88

Height 88

Body mass index 88

Absenteeism time in hours 88

dtype: int64

In [28]:



df[df['Disciplinary failure']**==**1].count()

Out[28]:

ID 40

Reason for absence 40

Month of absence 40

Day of the week 40

Seasons 40

Transportation expense 40

Distance from Residence to Work 40

Service time 40

Age 40

Work load Average/day 40

Hit target 40

Disciplinary failure 40

Education 40

Son 40

Social drinker 40

Social smoker 40

Pet 40

Weight 40

Height 40

Body mass index 40

Absenteeism time in hours 40

dtype: int64

In [29]:



roa **=** df.groupby('Reason for absence')

print(roa['Absenteeism time in hours'].max())

​

​

Reason for absence

0 0

1 64

2 24

3 8

4 8

5 8

6 120

7 40

8 8

9 112

10 48

11 120

12 112

13 120

14 40

15 8

16 3

17 8

18 80

19 104

21 8

22 16

23 16

24 8

25 8

26 16

27 3

28 24

Name: Absenteeism time in hours, dtype: int64

In [30]:



roa **=** df.groupby('Reason for absence')

roa['Absenteeism time in hours'].max()

Out[30]:

Reason for absence

0 0

1 64

2 24

3 8

4 8

5 8

6 120

7 40

8 8

9 112

10 48

11 120

12 112

13 120

14 40

15 8

16 3

17 8

18 80

19 104

21 8

22 16

23 16

24 8

25 8

26 16

27 3

28 24

Name: Absenteeism time in hours, dtype: int64

In [31]:



X **=** df.iloc[:, 1:20].values

y **=** df.iloc[:,20:].values.reshape(**-**1,1)

​

In [36]:



**from** sklearn.svm **import** SVR

​

In [38]:



**from** sklearn.model\_selection **import** train\_test\_split

X\_train, X\_test, y\_train, y\_test **=** train\_test\_split(X, y, test\_size **=** 0.5, random\_state **=** 0)

**from** sklearn.preprocessing **import** StandardScaler

sc\_X **=** StandardScaler()

sc\_y **=** StandardScaler()

X **=** sc\_X.fit\_transform(X)

y **=** sc\_y.fit\_transform(y)

prediction **=** sc\_y.inverse\_transform(svr\_regressor.predict(sc\_X.transform(X)))

X.shape

y.shape

Out[38]:

(740, 1)

In [43]:



y\_pred **=** svr\_regressor.predict(X\_test)

y\_pred **=** sc\_y.inverse\_transform(y\_pred)

**from** sklearn.metrics **import** mean\_squared\_error, explained\_variance\_score

print('MSE:{}'.format(mean\_squared\_error(y\_test,y\_pred)))

print('Explained variance score:{}'.format(explained\_variance\_score(y\_test,y\_pred)))

MSE:15.208462996904817

Explained variance score:-4.440892098500626e-16

In [51]:



plt.plot(X\_train,X\_test)

Out[51]:

[<matplotlib.lines.Line2D at 0x176fc3bfba8>,

<matplotlib.lines.Line2D at 0x176fc3bfcf8>,

<matplotlib.lines.Line2D at 0x176fc3bfe48>,

<matplotlib.lines.Line2D at 0x176fc3bff98>,

<matplotlib.lines.Line2D at 0x176fc3c8128>,

<matplotlib.lines.Line2D at 0x176fc3c8278>,

<matplotlib.lines.Line2D at 0x176fc3c83c8>,

<matplotlib.lines.Line2D at 0x176fc3c8518>,

<matplotlib.lines.Line2D at 0x176fc3c8668>,

<matplotlib.lines.Line2D at 0x176fc3c87b8>,

<matplotlib.lines.Line2D at 0x176fc391c18>,

<matplotlib.lines.Line2D at 0x176fc3c8a20>,

<matplotlib.lines.Line2D at 0x176fc3c8b70>,

<matplotlib.lines.Line2D at 0x176fc3c8cc0>,

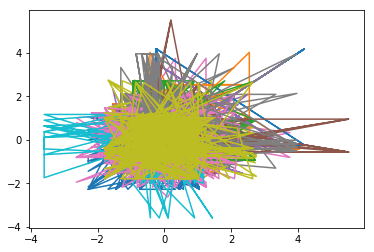
<matplotlib.lines.Line2D at 0x176fc3c8e10>,

<matplotlib.lines.Line2D at 0x176fc3c8f60>,

<matplotlib.lines.Line2D at 0x176fc3ce0f0>,

<matplotlib.lines.Line2D at 0x176fc3ce240>,

<matplotlib.lines.Line2D at 0x176fc3ce390>]



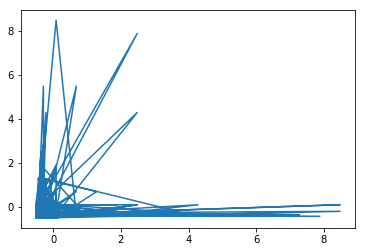
In [52]:



plt.plot(y\_train,y\_test)

Out[52]:

[<matplotlib.lines.Line2D at 0x176fd415860>]



In [53]:



plt.plot(X\_train,y\_train)

Out[53]:

[<matplotlib.lines.Line2D at 0x176fd462fd0>,

<matplotlib.lines.Line2D at 0x176fd46d160>,

<matplotlib.lines.Line2D at 0x176fd46d2b0>,

<matplotlib.lines.Line2D at 0x176fd46d400>,

<matplotlib.lines.Line2D at 0x176fd46d550>,

<matplotlib.lines.Line2D at 0x176fd46d6a0>,

<matplotlib.lines.Line2D at 0x176fd46d7f0>,

<matplotlib.lines.Line2D at 0x176fd46d940>,

<matplotlib.lines.Line2D at 0x176fd46da90>,

<matplotlib.lines.Line2D at 0x176fd46dbe0>,

<matplotlib.lines.Line2D at 0x176fd4440f0>,

<matplotlib.lines.Line2D at 0x176fd46de48>,

<matplotlib.lines.Line2D at 0x176fd46df98>,

<matplotlib.lines.Line2D at 0x176fd472128>,

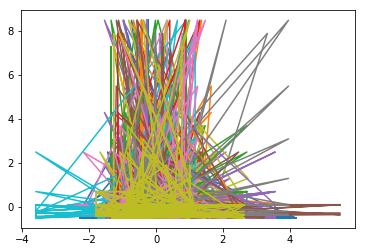
<matplotlib.lines.Line2D at 0x176fd472278>,

<matplotlib.lines.Line2D at 0x176fd4723c8>,

<matplotlib.lines.Line2D at 0x176fd472518>,

<matplotlib.lines.Line2D at 0x176fd472668>,

<matplotlib.lines.Line2D at 0x176fd4727b8>]



In [54]:



plt.plot(X\_test,y\_test)

Out[54]:

[<matplotlib.lines.Line2D at 0x176fd4e5ef0>,

<matplotlib.lines.Line2D at 0x176fd4f0080>,

<matplotlib.lines.Line2D at 0x176fd4f01d0>,

<matplotlib.lines.Line2D at 0x176fd4f0320>,

<matplotlib.lines.Line2D at 0x176fd4f0470>,

<matplotlib.lines.Line2D at 0x176fd4f05c0>,

<matplotlib.lines.Line2D at 0x176fd4f0710>,

<matplotlib.lines.Line2D at 0x176fd4f0860>,

<matplotlib.lines.Line2D at 0x176fd4f09b0>,

<matplotlib.lines.Line2D at 0x176fd4f0b00>,

<matplotlib.lines.Line2D at 0x176fd4b9390>,

<matplotlib.lines.Line2D at 0x176fd4f0d68>,

<matplotlib.lines.Line2D at 0x176fd4f0eb8>,

<matplotlib.lines.Line2D at 0x176fd4f4048>,

<matplotlib.lines.Line2D at 0x176fd4f4198>,

<matplotlib.lines.Line2D at 0x176fd4f42e8>,

<matplotlib.lines.Line2D at 0x176fd4f4438>,

<matplotlib.lines.Line2D at 0x176fd4f4588>,

<matplotlib.lines.Line2D at 0x176fd4f46d8>]

