Business Scenario

When deciding upon resourcing plans it is important to consider likely fault volumes. Since the orders that fuel the provision work are typically taken 2 weeks in advance, a forward looking forecast of fault volumes would enable a more accurate prediction of the number of provision volumes to take on each week.
You have been provided with the fault volume and rainfall data for a single geographic area for 3 years. Please use this data to create a model to predict the expected daily fault intake based on weather events. Detailed explanation of the fields provided is given below.
At the interview you will be given 10 minutes to present your findings. The presentation should give an overview of your findings and be targeted at a senior stakeholder level. Please provide a copy of your findings for review before the interview.

Things you may want to consider in the model;
1 Both short term and long term impact of rainfall
2 Impact of weather events on different fault types
3 Other things that might impact fault volume

Data provided:
Fault Data: This contains information of individual faults for a single region over a 4 year period.
Fault ID: Fault identifier

Report Date: The date that the fault was reported
Initial MFL: Location of the fault;
CA: Customer appointed (home)
CE: D-side overhead network
EX: Exchange
FU: Frames (exchange)
LN: E-side Underground network
OK: Line has tested OK
OTHER: Other faults
Rainfall: This contains daily rainfall information for the same 4 year period
Observation Date: Date
Rainfall mm: Rainfall (in mm)
Calendar Lookup: This contains a lookup to give calendar information such as day of week and bank holidays
Actual Date: Date
Day of Week: Day of Week
Day Num Cal Week: Numeric value of day within the week
Day Num Cal Month: Numeric value of day within the month
Day Num Cal Year: Numeric value of day within the year
Bank holiday: Flags the bank holiday (Y: bank holiday)

In [544]:

```
# Import Libraries
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
import pandas as pd
import numpy as np
import os
import matplotlib.pyplot as plt
import matplotlib.dates as mdates
import seaborn as sb
import datetime
from datetime import timedelta,datetime
from sklearn import metrics
```

In [692]:

```
X = pd.read_excel('RainFallData.xlsx',index_col = 0)
Y = pd.read_excel('FaultData.xlsx', index_col = 0)
```

In [693]:

```
X['OBS_DATE'] = X.index
print(X)
```

```
                 3.3 2014-01-16
2014-01-16
2014-01-17       2.0 2014-01-17
2014-01-18       9.2 2014-01-18
2014-01-19       2.6 2014-01-19
2014-01-20       0.2 2014-01-20
2014-01-21       5.6 2014-01-21
2014-01-22       2.3 2014-01-22
2014-01-23       1.5 2014-01-23
2014-01-24       6.4 2014-01-24
2014-01-25       3.2 2014-01-25
2014-01-26       9.9 2014-01-26
2014-01-27      10.0 2014-01-27
2014-01-28       8.0 2014-01-28
2014-01-29       2.1 2014-01-29
2014-01-30       0.0 2014-01-30
...              ...        ...
2017-02-28      11.5 2017-02-28
2017-03-01      14.4 2017-03-01
2017-03-02       1.9 2017-03-02
```

In [694]:

```
# Rainfall Every 1 to 7 Days
X['1_Day_Rain'] = [X['RAINFALL_MM'][max(0,i-1)] for i in range(0,len(X))]
X['2_Day_Rain'] = [X['RAINFALL_MM'][max(0,i-2)] for i in range(0,len(X))]
X['3_Day_Rain'] = [X['RAINFALL_MM'][max(0,i-3)] for i in range(0,len(X))]
X['4_Day_Rain'] = [X['RAINFALL_MM'][max(0,i-4)] for i in range(0,len(X))]
X['5_Day_Rain'] = [X['RAINFALL_MM'][max(0,i-5)] for i in range(0,len(X))]
X['6_Day_Rain'] = [X['RAINFALL_MM'][max(0,i-6)] for i in range(0,len(X))]
X['7_Day_Rain'] = [X['RAINFALL_MM'][max(0,i-7)] for i in range(0,len(X))]

## Cummulative Rain for each week
X['Cummulative_Rain'] = [sum(X['RAINFALL_MM'][max(0,i-7):i]) for i in range(0,len(X))]
X['YMD'] = [X['OBS_DATE'][i].year*10000+X['OBS_DATE'][i].month*100+X['OBS_DATE'][i].da
X
```

Out[694]:

OBSERVATION_DATE	RAINFALL_MM	OBS_DATE	1_Day_Rain	2_Day_Rain	3_Day_Rain	4_
2014-01-01	10.6	2014-01-01	10.6	10.6	10.6	
2014-01-02	4.2	2014-01-02	10.6	10.6	10.6	
2014-01-03	5.8	2014-01-03	4.2	10.6	10.6	
2014-01-04	1.8	2014-01-04	5.8	4.2	10.6	
2014-01-05	11.0	2014-01-05	1.8	5.8	4.2	
2014-01-06	5.6	2014-01-06	11.0	1.8	5.8	
2014-01-07	5.9	2014-01-07	5.6	11.0	1.8	
2014-01-08	14.8	2014-01-08	5.9	5.6	11.0	
2014-01-09	0.8	2014-01-09	14.8	5.9	5.6	
2014-01-10	5.2	2014-01-10	0.8	14.8	5.9	
2014-01-11	0.2	2014-01-11	5.2	0.8	14.8	
2014-01-12	3.8	2014-01-12	0.2	5.2	0.8	
2014-01-13	8.1	2014-01-13	3.8	0.2	5.2	
2014-01-14	7.6	2014-01-14	8.1	3.8	0.2	
2014-01-15	5.3	2014-01-15	7.6	8.1	3.8	
2014-01-16	3.3	2014-01-16	5.3	7.6	8.1	
2014-01-17	2.0	2014-01-17	3.3	5.3	7.6	
2014-01-18	9.2	2014-01-18	2.0	3.3	5.3	
2014-01-19	2.6	2014-01-19	9.2	2.0	3.3	
2014-01-20	0.2	2014-01-20	2.6	9.2	2.0	
2014-01-21	5.6	2014-01-21	0.2	2.6	9.2	
2014-01-22	2.3	2014-01-22	5.6	0.2	2.6	
2014-01-23	1.5	2014-01-23	2.3	5.6	0.2	
2014-01-24	6.4	2014-01-24	1.5	2.3	5.6	

OBSERVATION_DATE	RAINFALL_MM	OBS_DATE	1_Day_Rain	2_Day_Rain	3_Day_Rain	4_
2014-01-25	3.2	2014-01-25	6.4	1.5	2.3	
2014-01-26	9.9	2014-01-26	3.2	6.4	1.5	
2014-01-27	10.0	2014-01-27	9.9	3.2	6.4	
2014-01-28	8.0	2014-01-28	10.0	9.9	3.2	
2014-01-29	2.1	2014-01-29	8.0	10.0	9.9	
2014-01-30	0.0	2014-01-30	2.1	8.0	10.0	
...	...	...	...	...	...	
2017-02-28	11.5	2017-02-28	7.7	11.0	11.6	
2017-03-01	14.4	2017-03-01	11.5	7.7	11.0	
2017-03-02	1.9	2017-03-02	14.4	11.5	7.7	
2017-03-03	5.3	2017-03-03	1.9	14.4	11.5	
2017-03-04	3.8	2017-03-04	5.3	1.9	14.4	
2017-03-05	11.4	2017-03-05	3.8	5.3	1.9	
2017-03-06	4.3	2017-03-06	11.4	3.8	5.3	
2017-03-07	7.8	2017-03-07	4.3	11.4	3.8	
2017-03-08	0.1	2017-03-08	7.8	4.3	11.4	
2017-03-09	0.9	2017-03-09	0.1	7.8	4.3	
2017-03-10	0.8	2017-03-10	0.9	0.1	7.8	
2017-03-11	1.7	2017-03-11	0.8	0.9	0.1	
2017-03-12	0.8	2017-03-12	1.7	0.8	0.9	
2017-03-13	0.1	2017-03-13	0.8	1.7	0.8	
2017-03-14	0.0	2017-03-14	0.1	0.8	1.7	
2017-03-15	0.3	2017-03-15	0.0	0.1	0.8	
2017-03-16	1.8	2017-03-16	0.3	0.0	0.1	
2017-03-17	10.9	2017-03-17	1.8	0.3	0.0	
2017-03-18	2.6	2017-03-18	10.9	1.8	0.3	
2017-03-19	5.1	2017-03-19	2.6	10.9	1.8	
2017-03-21	8.4	2017-03-21	5.1	2.6	10.9	
2017-03-22	2.2	2017-03-22	8.4	5.1	2.6	
2017-03-23	0.1	2017-03-23	2.2	8.4	5.1	
2017-03-24	0.0	2017-03-24	0.1	2.2	8.4	
2017-03-25	0.0	2017-03-25	0.0	0.1	2.2	
2017-03-26	0.0	2017-03-26	0.0	0.0	0.1	
2017-03-27	0.0	2017-03-27	0.0	0.0	0.0	
2017-03-28	4.9	2017-03-28	0.0	0.0	0.0	
2017-03-29	9.0	2017-03-29	4.9	0.0	0.0	
2017-03-30	0.0	2017-03-30	9.0	4.9	0.0	

In [695]:

```
print(type(X))
print(type(Y))
```

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

Calculate Different types of Faults for 7, 14 and 21 Days

In [696]:

```
Y.head()
(Y['REPORT_DATE'][0]).day
```

Out[696]:

```
1
```

In [697]:

```
Y['YMD'] = [Y['REPORT_DATE'][i].year*10000+Y['REPORT_DATE'][i].month*100+Y['REPORT_DATE
Y.head()
```

Out[697]:

Fault ID	REPORT_DATE	INITIAL_MFL	YMD
ID000001	2014-01-01	CE	20140101
ID000002	2014-01-01	CA	20140101
ID000003	2014-01-01	OTHER	20140101
ID000004	2014-01-01	OK	20140101
ID000005	2014-01-01	LN	20140101

In [698]:

```
Y['COUNT'] = 1
Z = pd.pivot_table(Y,values=['COUNT'],index='YMD',columns='INITIAL_MFL',aggfunc=sum).r
```

In [699]:

```
Z.columns=['YMD','FAULTS_CA','FAULTS_CE','FAULTS_EX','FAULTS_FU','FAULTS_LN','FAULTS_O
Z.head()
```

Out[699]:

	YMD	FAULTS_CA	FAULTS_CE	FAULTS_EX	FAULTS_FU	FAULTS_LN	FAULTS_OK	FA
0	20140101	15.0	40.0	6.0	4.0	38.0	26.0	
1	20140102	47.0	100.0	3.0	13.0	67.0	83.0	
2	20140103	63.0	87.0	6.0	14.0	77.0	79.0	
3	20140104	24.0	60.0	3.0	6.0	34.0	47.0	
4	20140105	16.0	28.0	2.0	4.0	31.0	41.0	

In [700]:

```
for fault in Z.columns[1:]:
    newfault = fault.replace('FAULTS_','15to21Faults_')
    Z[fault]=Z[fault].fillna(0).replace(np.nan,0)
    Z[newfault]=[sum(Z[fault][min(i+14,len(Z)):min(i+21,len(Z))]) for i in range(0,len
```

In [701]:

```
Z.tail(50)
```

1171	20170317	31.0	84.0	0.0	17.0	45.0	93.0	
1172	20170318	13.0	59.0	0.0	11.0	22.0	52.0	
1173	20170319	19.0	53.0	0.0	11.0	16.0	42.0	
1174	20170320	35.0	102.0	0.0	12.0	56.0	114.0	
1175	20170321	23.0	90.0	7.0	15.0	49.0	95.0	
1176	20170322	27.0	98.0	1.0	14.0	42.0	80.0	
1177	20170323	32.0	116.0	0.0	12.0	47.0	92.0	
1178	20170324	29.0	98.0	2.0	10.0	38.0	99.0	
1179	20170325	17.0	56.0	0.0	7.0	16.0	55.0	
1180	20170326	9.0	37.0	0.0	5.0	12.0	34.0	
1181	20170327	22.0	100.0	2.0	11.0	29.0	87.0	
1182	20170328	26.0	81.0	1.0	9.0	20.0	88.0	

In [702]:

```
## Merge DataFrame
df=X.merge(Z,on='YMD')
df = df[7:len(df)-21]
df = df.reset_index()
df
```

...	...	...	...	...	...	...	...	...
1127	1134	1.5	2017-02-07	6.7	0.1	4.2	10.6	4.3
1128	1135	0.0	2017-02-08	1.5	6.7	0.1	4.2	10.6
1129	1136	0.0	2017-02-09	0.0	1.5	6.7	0.1	4.2
1130	1137	0.0	2017-02-10	0.0	0.0	1.5	6.7	0.1
1131	1138	1.1	2017-02-11	0.0	0.0	0.0	1.5	6.7
1132	1139	0.3	2017-02-12	1.1	0.0	0.0	0.0	1.5
1133	1140	0.0	2017-02-13	0.3	1.1	0.0	0.0	0.0
1134	1141	0.2	2017-02-14	0.0	0.3	1.1	0.0	0.0
1135	1142	0.4	2017-02-15	0.2	0.0	0.3	1.1	0.0

In [703]:

```
train, test = train_test_split(df, test_size=0.2)
```

In [704]:

```
train.head()
```

Out[704]:

	index	RAINFALL_MM	OBS_DATE	1_Day_Rain	2_Day_Rain	3_Day_Rain	4_Day_Rain	5_I
860	867	0.0	2016-05-16	0.0	0.0	0.0	0.0	
840	847	2.0	2016-04-26	1.8	0.6	0.1	0.0	
987	994	0.0	2016-09-20	0.0	3.9	0.0	0.1	
452	459	0.0	2015-04-04	1.6	3.7	7.3	1.3	
220	227	1.3	2014-08-16	0.5	1.3	1.7	1.8	

5 rows × 26 columns

In [705]:

```
1 test.head()
```

Out[705]:

	index	RAINFALL_MM	OBS_DATE	1_Day_Rain	2_Day_Rain	3_Day_Rain	4_Day_Rain	5
985	992	3.9	2016-09-18	0.0	0.1	0.2	0.0	
821	828	4.6	2016-04-07	8.2	2.4	3.2	11.4	
966	973	2.3	2016-08-30	0.0	0.1	6.5	0.1	
1132	1139	0.3	2017-02-12	1.1	0.0	0.0	0.0	
67	74	0.2	2014-03-16	0.0	0.0	0.2	0.1	

5 rows × 26 columns

CA - FAULT MODELS

Linear Regression

In [931]:

```
1 # Split Train Test Data for CA Faults
2 df_ca = df[['Cummulative_Rain','15to21Faults_CA']]
3 df_ca_train, df_ca_test = train_test_split(df_ca, test_size=0.2,random_state=1801)
```

In [950]:

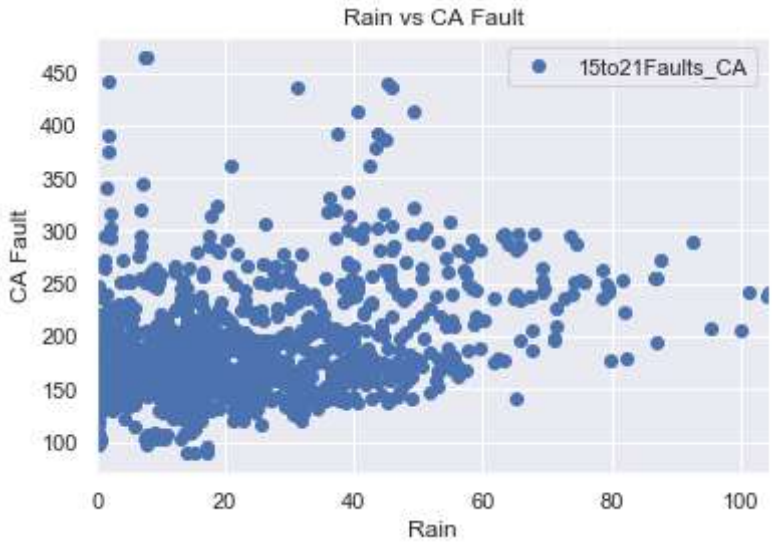
```
1 df_ca.corr()
```

Out[950]:

	Cummulative_Rain	15to21Faults_CA
Cummulative_Rain	1.000000	0.293605
15to21Faults_CA	0.293605	1.000000

In [932]:

```
df_ca.plot(x='Cummulative_Rain', y='15to21Faults_CA', style='o')
plt.title('Rain vs CA Fault')
plt.xlabel('Rain')
plt.ylabel('CA Fault')
plt.show()
```

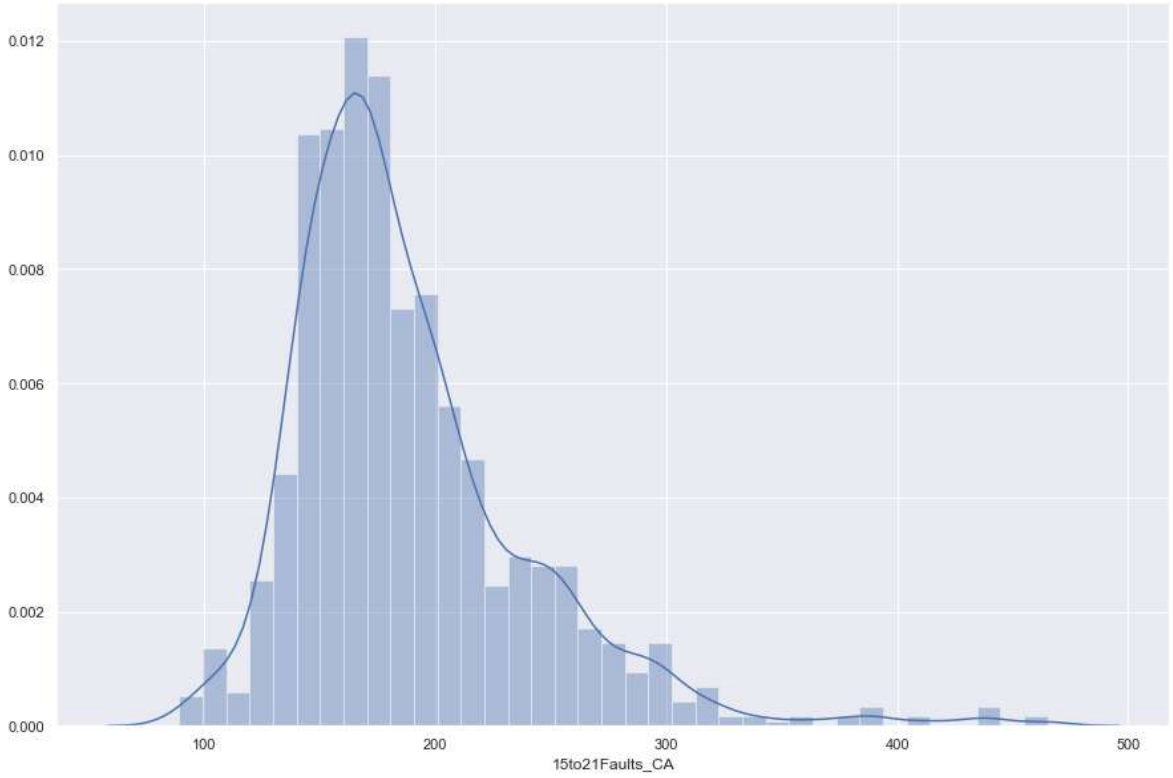


In [933]:

```
# Check Average Faults
plt.figure(figsize=(15,10))
plt.tight_layout()
sb.distplot(df_ca['15to21Faults_CA'])
```

Out[933]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x15c0562a9e8>
```



In [934]:

```
# Find TRAIN Set Data for 15 to 21 Days for CA Faults
df_ca_train_np_rain=df_ca_train['Cummulative_Rain'].to_numpy()
df_ca_train_np_fault = df_ca_train['15to21Faults_CA'].to_numpy()

# Reshape test train data
df_ca_train_rain = df_ca_train_np_rain.reshape(len(df_ca_train_np_rain),1)
df_ca_train_fault = df_ca_train_np_fault.reshape(len(df_ca_train_np_fault),1)
```

In [935]:

```
lr = LinearRegression().fit(df_ca_train_rain,df_ca_train_fault)
```

In [936]:

```
print(lr.coef_)
print(lr.intercept_)
```

```
[[0.778806]]
[170.60607576]
```

In [937]:

```
# Find TEST Set Data for 15 to 21 Days for CA Faults
df_ca_test_np_rain=df_ca_test['Cummulative_Rain'].to_numpy()
df_ca_test_np_fault = df_ca_test['15to21Faults_CA'].to_numpy()

df_ca_test_rain = df_ca_test_np_rain.reshape(len(df_ca_test_np_rain),1)
df_ca_test_fault = df_ca_test_np_fault.reshape(len(df_ca_test_np_fault),1)
```

In [938]:

```
# Train Model Accuracy
lr.score(df_ca_train_rain, df_ca_train_fault)
```

Out[938]:

```
0.07873837961092411
```

In [939]:

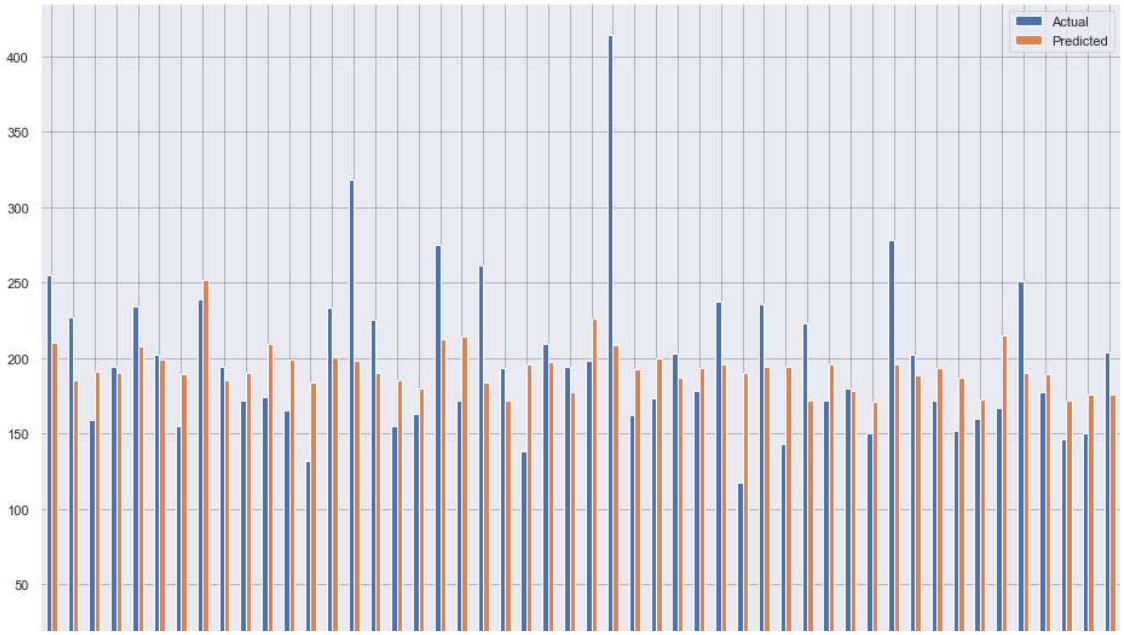
```
# testing model accuracy in test set
lr.score(df_ca_test_rain, df_ca_test_fault)
```

Out[939]:

```
0.11870411713873263
```

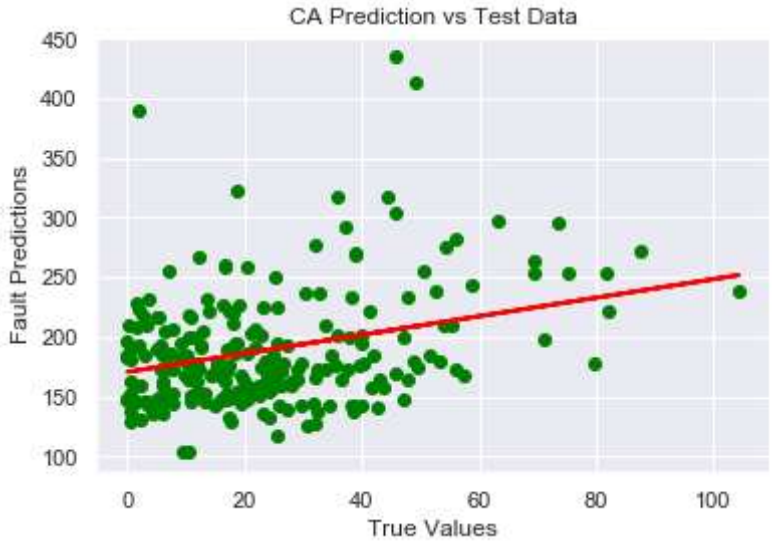
In [940]:

```
predict_linear = lr.predict(df_ca_test_rain)
actual_pred_linear= pd.DataFrame({'Actual': df_ca_test_fault.flatten(), 'Predicted': p
bar_plot_df = actual_pred_linear.head(50)
bar_plot_df.plot(kind='bar',figsize=(16,10))
plt.grid(which='major', linestyle='-', linewidth='0.5', color='grey')
plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')
plt.show()
```



In [949]:

```
plt.scatter(df_ca_test_rain, df_ca_test_fault,  color='green')
plt.plot(df_ca_test_rain, predict_linear, color='red', linewidth=2)
plt.title('CA Prediction vs Test Data')
plt.xlabel('True Values')
plt.ylabel('Fault Predictions')
plt.show()
```



In [728]:

```
rmspe = (np.sqrt(np.mean(np.square((df_ca_test_fault - predict_linear) / df_ca_test_fa
print('Root Mean Squared Error Percentage for CA Faults: ', rmspe,'%')
```

```
Root Mean Squared Error Percentage for CA Faults:  22.58179783574731 %
```

OK FAULT MODELS

In [729]:

```
df_ok = df[['Cummulative_Rain','15to21Faults_OK']]
df_ok_train, df_ok_test = train_test_split(df_ok, test_size=0.2,random_state=1801)
```

In [951]:

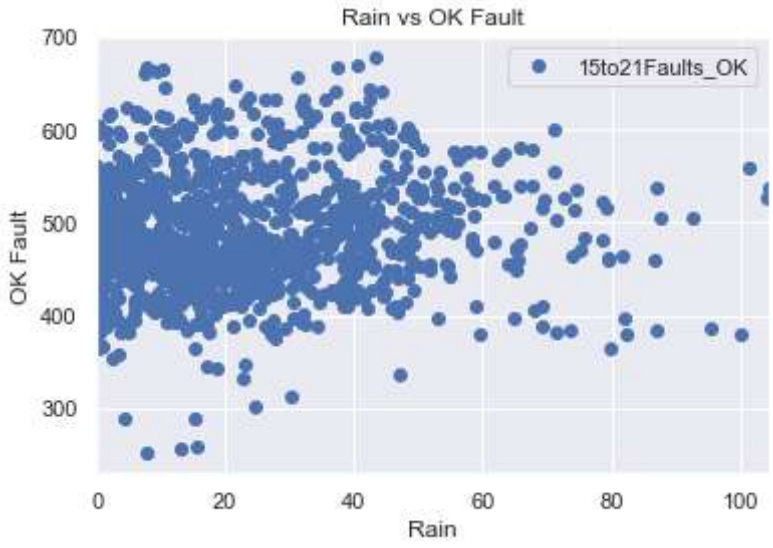
```
df_ok.corr()
```

Out[951]:

	Cummulative_Rain	15to21Faults_OK
Cummulative_Rain	1.000000	0.109644
15to21Faults_OK	0.109644	1.000000

In [730]:

```
df_ok.plot(x='Cummulative_Rain', y='15to21Faults_OK', style='o')
plt.title('Rain vs OK Fault')
plt.xlabel('Rain')
plt.ylabel('OK Fault')
plt.show()
```

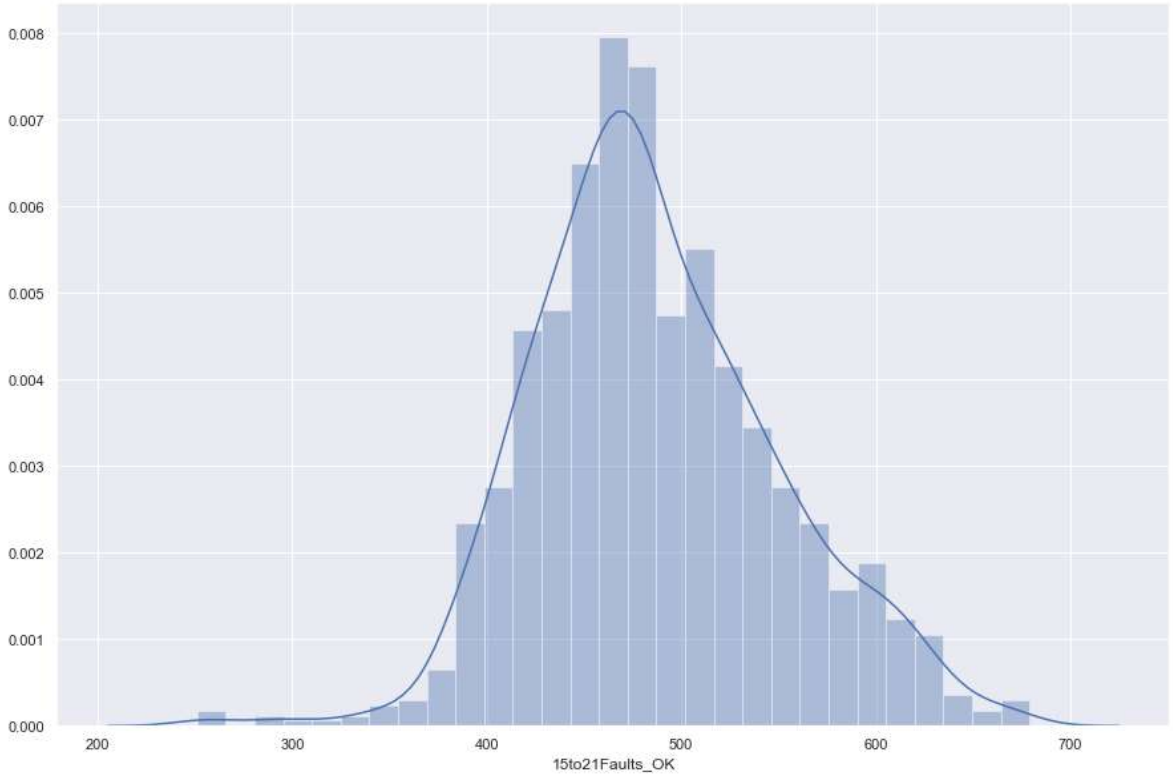


In [731]:

```
# Check Average Faults
plt.figure(figsize=(15,10))
plt.tight_layout()
sb.distplot(df_ok['15to21Faults_OK'])
```

Out[731]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x15c07143a20>
```



In [732]:

```
# Find TRAIN Set Data for 15 to 21 Days for OK Faults
df_ok_train_np_rain=df_ok_train['Cummulative_Rain'].to_numpy()
df_ok_train_np_fault = df_ok_train['15to21Faults_OK'].to_numpy()

# Reshape test train data
df_ok_train_rain = df_ok_train_np_rain.reshape(len(df_ok_train_np_rain),1)
df_ok_train_fault = df_ok_train_np_fault.reshape(len(df_ok_train_np_fault),1)
```

In [733]:

```
lr = LinearRegression().fit(df_ok_train_rain,df_ok_train_fault)
```

In [734]:

```
# Train Model Accuracy
lr.score(df_ok_train_rain, df_ok_train_fault)
```

Out[734]:

```
0.022303093262508348
```

In [735]:

```
# Find TEST Set Data for 15 to 21 Days for OK Faults
df_ok_test_np_rain=df_ok_test['Cummulative_Rain'].to_numpy()
df_ok_test_np_fault = df_ok_test['15to21Faults_OK'].to_numpy()

df_ok_test_rain = df_ok_test_np_rain.reshape(len(df_ok_test_np_rain),1)
df_ok_test_fault = df_ok_test_np_fault.reshape(len(df_ok_test_np_fault),1)
```

In [736]:

```
# testing model accuracy in test set
lr.score(df_ok_test_rain, df_ok_test_fault)
```

Out[736]:

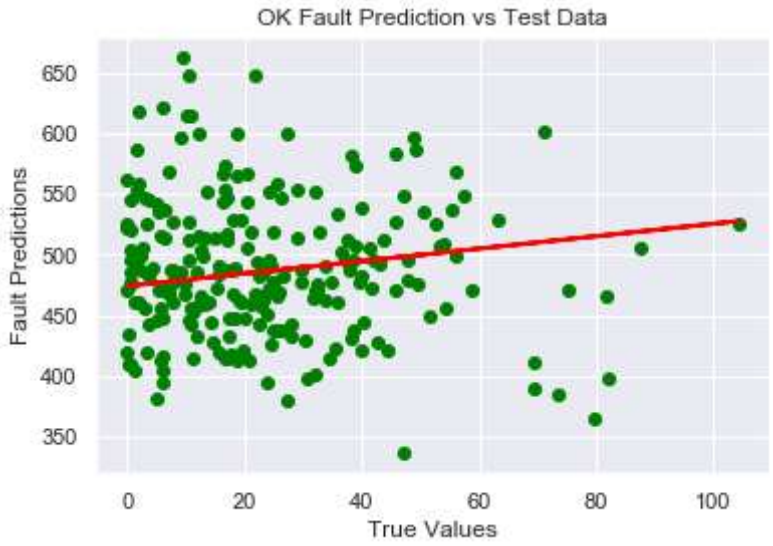
```
-0.05419377300430939
```

In [737]:

```
pred_OK_Fault_linear = lr.predict(df_ok_test_rain)
OK_actual_pred_linear= pd.DataFrame({'Actual': df_ok_test_fault.flatten(), 'Predicted'
bar_plot_df = OK_actual_pred_linear.head(50)
bar_plot_df.plot(kind='bar',figsize=(16,10))
plt.grid(which='major', linestyle='-', linewidth='0.5', color='green')
plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')
plt.show()
```

In [943]:

```
plt.scatter(df_ok_test_rain, df_ok_test_fault,  color='green')
plt.plot(df_ok_test_rain, pred_OK_Fault_linear, color='red', linewidth=2)
plt.title('OK Fault Prediction vs Test Data')
plt.xlabel('True Values')
plt.ylabel('Fault Predictions')
plt.show()
```



In [808]:

```
rmspe_ok = (np.sqrt(np.mean(np.square((df_ok_test_fault - pred_OK_Fault_linear) / df_o
print('Root Mean Squared Error Percentage for OK Faults: ', rmspe_ok,'%')
```

```
Root Mean Squared Error Percentage for OK Faults:  12.093466744257546 %
```

CE FAULT MODEL

Linear Regression on Faults CE

In [740]:

```
df_ce = df[['Cummulative_Rain','15to21Faults_CE']]
df_ce_train, df_ce_test = train_test_split(df_ce, test_size=0.2,random_state=1801)
```

In [952]:

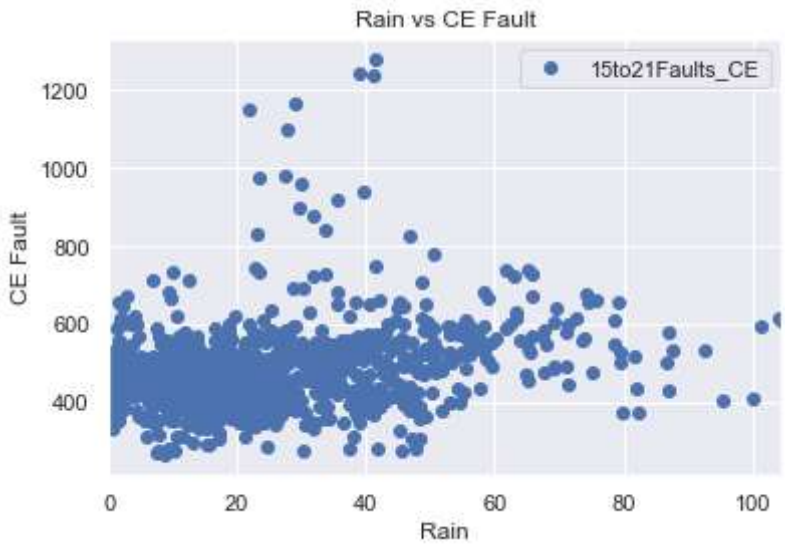
```
df_ce.corr()
```

Out[952]:

	Cummulative_Rain	15to21Faults_CE
Cummulative_Rain	1.000000	0.248735
15to21Faults_CE	0.248735	1.000000

In [741]:

```
df_ce.plot(x='Cummulative_Rain', y='15to21Faults_CE', style='o')
plt.title('Rain vs CE Fault')
plt.xlabel('Rain')
plt.ylabel('CE Fault')
plt.show()
```

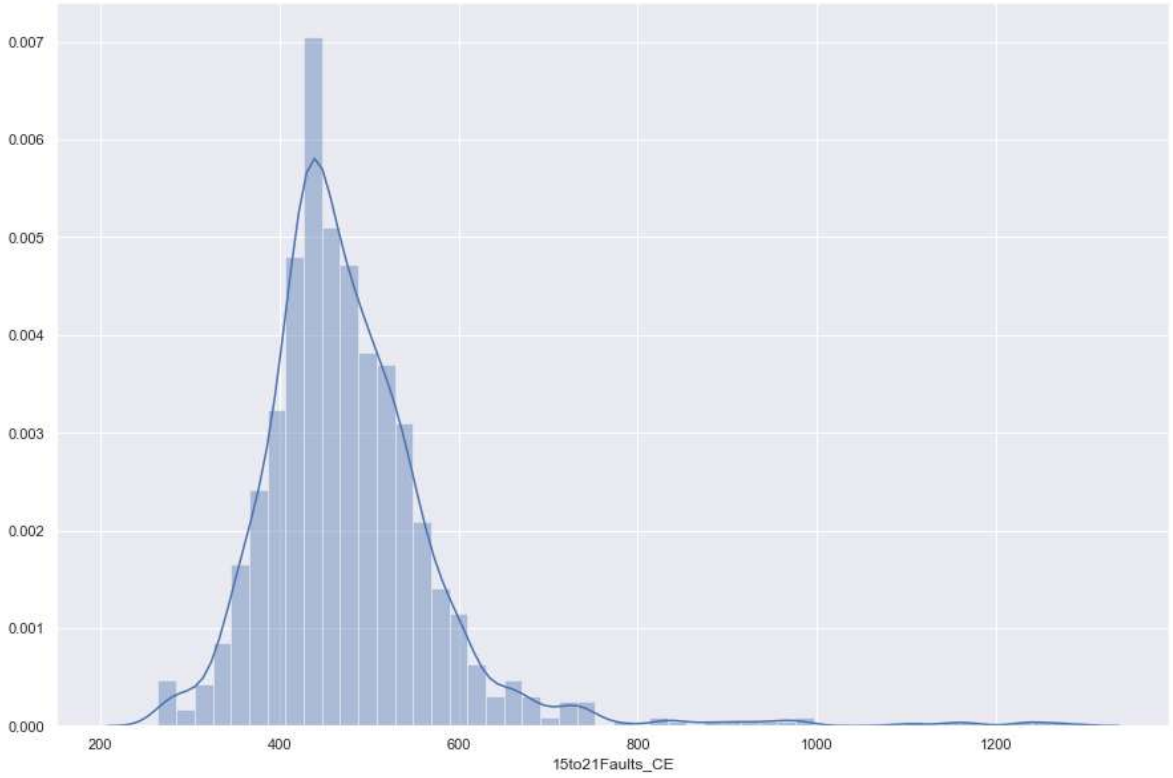


In [742]:

```
# Check Average Faults
plt.figure(figsize=(15,10))
plt.tight_layout()
sb.distplot(df_ce['15to21Faults_CE'])
```

Out[742]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x15c0bfa9ba8>
```



In [743]:

```
# Find TRAIN Set Data for 15 to 21 Days for CA Faults
df_ce_train_np_rain=df_ce_train['Cummulative_Rain'].to_numpy()
df_ce_train_np_fault = df_ce_train['15to21Faults_CE'].to_numpy()

# Reshape test train data
df_ce_train_rain = df_ce_train_np_rain.reshape(len(df_ce_train_np_rain),1)
df_ce_train_fault = df_ce_train_np_fault.reshape(len(df_ce_train_np_fault),1)
```

In [744]:

```
lr = LinearRegression().fit(df_ce_train_rain,df_ce_train_fault)
```

In [745]:

```
lr.score(df_ce_train_rain, df_ce_train_fault)
```

Out[745]:

```
0.06342011283905269
```

In [746]:

```
# Find TEST Set Data for 15 to 21 Days for CA Faults
df_ce_test_np_rain=df_ce_test['Cummulative_Rain'].to_numpy()
df_ce_test_np_fault = df_ce_test['15to21Faults_CE'].to_numpy()

df_ce_test_rain = df_ce_test_np_rain.reshape(len(df_ce_test_np_rain),1)
df_ce_test_fault = df_ce_test_np_fault.reshape(len(df_ce_test_np_fault),1)
```

In [747]:

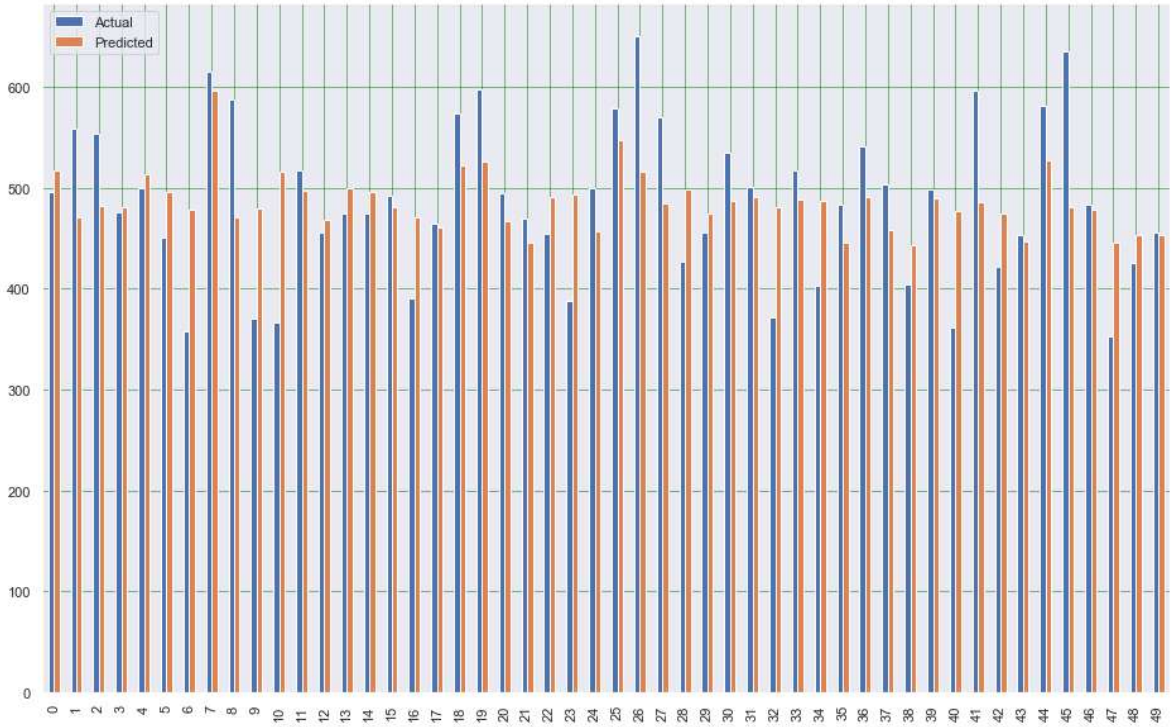
```
# testing model accuracy in test set
lr.score(df_ce_test_rain, df_ce_test_fault)
```

Out[747]:

```
0.04189087198405883
```

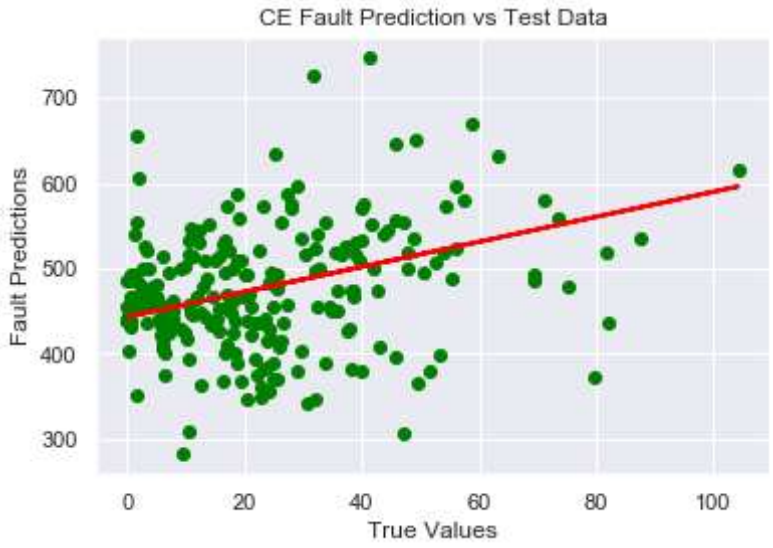
In [748]:

```
pred_ce_Fault_linear = lr.predict(df_ce_test_rain)
ce_actual_pred_linear= pd.DataFrame({'Actual': df_ce_test_fault.flatten(), 'Predicted'
bar_plot_df = ce_actual_pred_linear.head(50)
bar_plot_df.plot(kind='bar',figsize=(16,10))
plt.grid(which='major', linestyle='-', linewidth='0.5', color='green')
plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')
plt.show()
```



In [944]:

```
plt.scatter(df_ce_test_rain, df_ce_test_fault,  color='green')
plt.plot(df_ce_test_rain, pred_ce_Fault_linear, color='red', linewidth=2)
plt.title('CE Fault Prediction vs Test Data')
plt.xlabel('True Values')
plt.ylabel('Fault Predictions')
plt.show()
```



In [750]:

```
rmspe_ce = (np.sqrt(np.mean(np.square((df_ce_test_fault - pred_ce_Fault_linear) / df_c
print('Root Mean Squared Error Percentage for CE Faults: ', rmspe_ok,'%')
```

```
Root Mean Squared Error Percentage for CE Faults:  12.093466744257546 %
```

EX FAULT MODEL

Linear Regression

In [751]:

```
df_ex = df[['Cummulative_Rain','15to21Faults_EX']]
df_ex_train, df_ex_test = train_test_split(df_ex, test_size=0.2,random_state=1801)
```

In [953]:

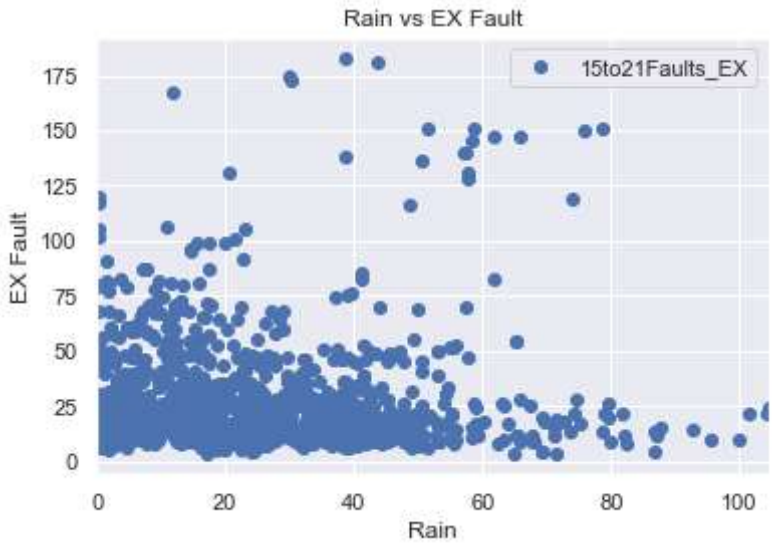
```
1 df_ex.corr()
```

Out[953]:

	Cummulative_Rain	15to21Faults_EX
Cummulative_Rain	1.000000	0.014196
15to21Faults_EX	0.014196	1.000000

In [752]:

```
1 df_ex.plot(x='Cummulative_Rain', y='15to21Faults_EX', style='o')
2 plt.title('Rain vs EX Fault')
3 plt.xlabel('Rain')
4 plt.ylabel('EX Fault')
5 plt.show()
```

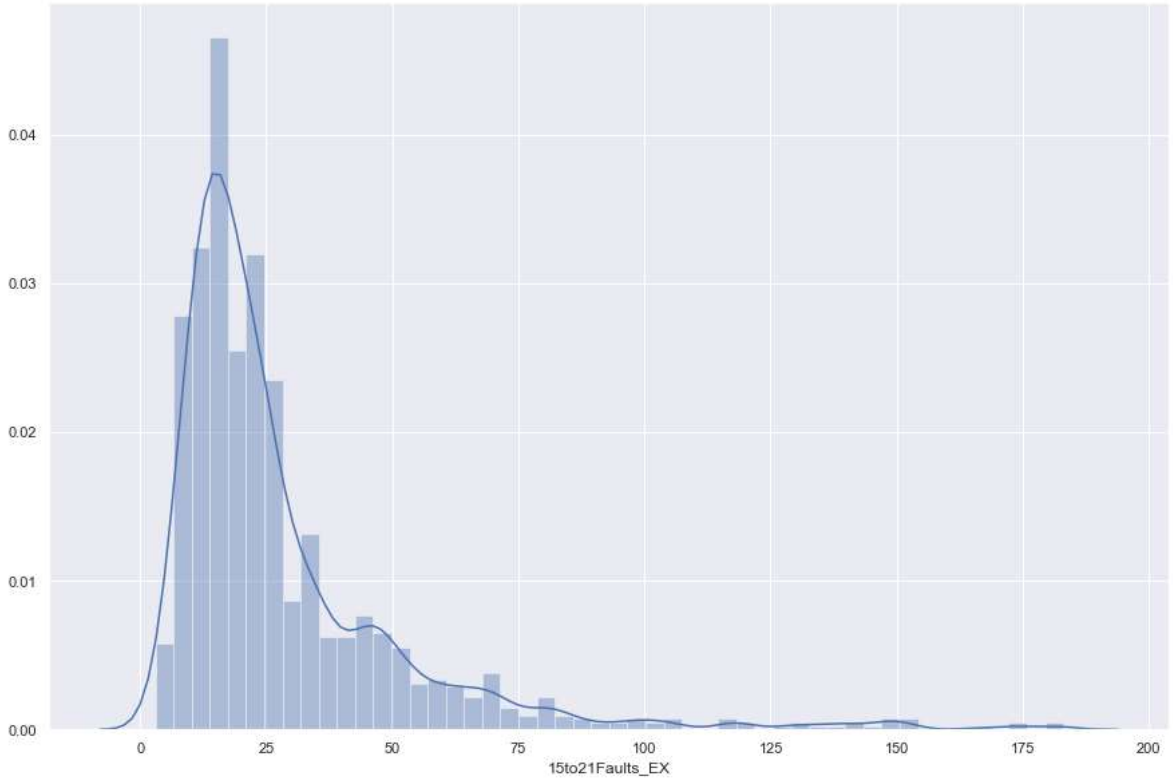


In [753]:

```
# Check Average Faults
plt.figure(figsize=(15,10))
plt.tight_layout()
sb.distplot(df_ex['15to21Faults_EX'])
```

Out[753]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x15c052f7208>
```



In [754]:

```
# Find TRAIN Set Data for 15 to 21 Days for CA Faults
df_ex_train_np_rain=df_ex_train['Cummulative_Rain'].to_numpy()
df_ex_train_np_fault = df_ex_train['15to21Faults_EX'].to_numpy()

# Reshape test train data
df_ex_train_rain = df_ex_train_np_rain.reshape(len(df_ex_train_np_rain),1)
df_ex_train_fault = df_ex_train_np_fault.reshape(len(df_ex_train_np_fault),1)
```

In [755]:

```
lr = LinearRegression().fit(df_ex_train_rain,df_ex_train_fault)
```

In [756]:

```
lr.score(df_ex_train_rain, df_ex_train_fault)
```

Out[756]:

```
0.0009182826737963444
```

In [757]:

```
# Find TEST Set Data for 15 to 21 Days for CA Faults
df_ex_test_np_rain=df_ex_test['Cummulative_Rain'].to_numpy()
df_ex_test_np_fault = df_ex_test['15to21Faults_EX'].to_numpy()

df_ex_test_rain = df_ex_test_np_rain.reshape(len(df_ex_test_np_rain),1)
df_ex_test_fault = df_ex_test_np_fault.reshape(len(df_ex_test_np_fault),1)
```

In [758]:

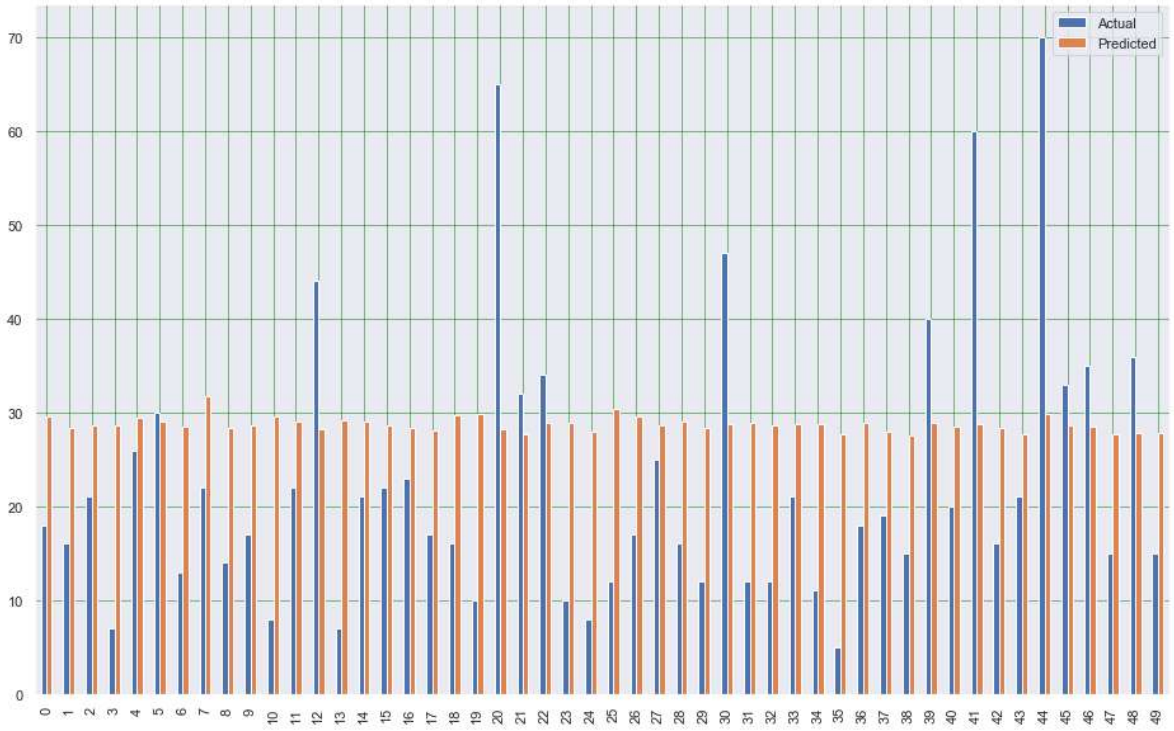
```
# testing model accuracy in test set
lr.score(df_ex_test_rain, df_ex_test_fault)
```

Out[758]:

```
-0.00497290884903534
```

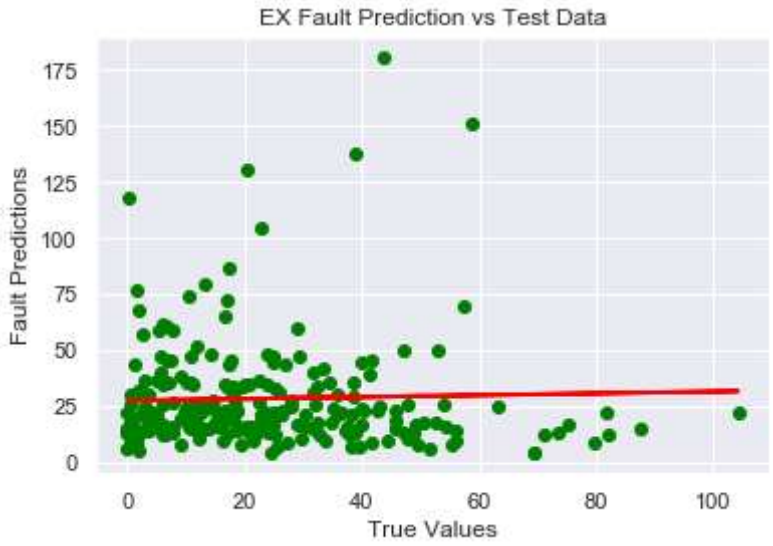
In [759]:

```
pred_ex_Fault_linear = lr.predict(df_ex_test_rain)
ex_actual_pred_linear= pd.DataFrame({'Actual': df_ex_test_fault.flatten(), 'Predicted'
bar_plot_df = ex_actual_pred_linear.head(50)
bar_plot_df.plot(kind='bar',figsize=(16,10))
plt.grid(which='major', linestyle='-', linewidth='0.5', color='green')
plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')
plt.show()
```



In [945]:

```
plt.scatter(df_ex_test_rain, df_ex_test_fault,  color='green')
plt.plot(df_ex_test_rain, pred_ex_Fault_linear, color='red', linewidth=2)
plt.title('EX Fault Prediction vs Test Data')
plt.xlabel('True Values')
plt.ylabel('Fault Predictions')
plt.show()
```



In [761]:

```
rmspe_ex = (np.sqrt(np.mean(np.square((df_ex_test_fault - pred_ex_Fault_linear) / df_e
print('Root Mean Squared Error Percentage for EX Faults: ', rmspe_ex,'%')
```

```
Root Mean Squared Error Percentage for EX Faults:  125.6560203361075 %
```

FU FAULT MODEL

Linear Regression

In [762]:

```
df_fu = df[['Cummulative_Rain','15to21Faults_FU']]
df_fu_train, df_fu_test = train_test_split(df_fu, test_size=0.2,random_state=1801)
```

In [954]:

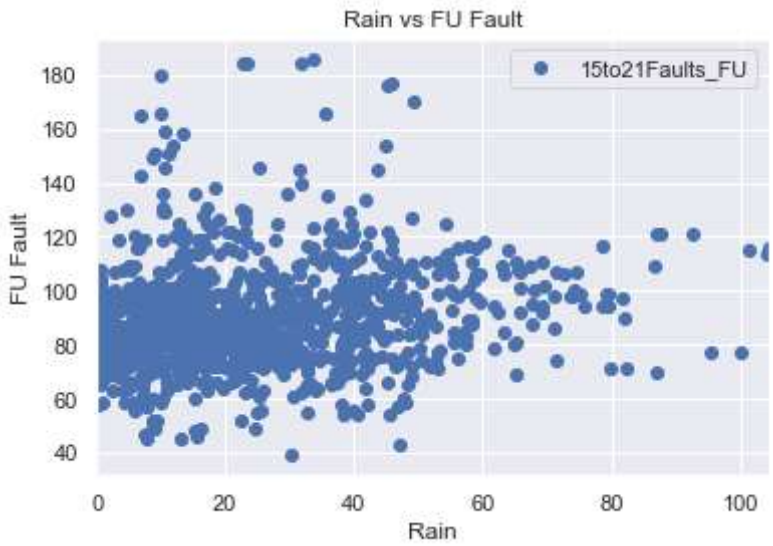
```
1 df_fu.corr()
```

Out[954]:

	Cummulative_Rain	15to21Faults_FU
Cummulative_Rain	1.000000	0.157248
15to21Faults_FU	0.157248	1.000000

In [763]:

```
1 df_fu.plot(x='Cummulative_Rain', y='15to21Faults_FU', style='o')
2 plt.title('Rain vs FU Fault')
3 plt.xlabel('Rain')
4 plt.ylabel('FU Fault')
5 plt.show()
```

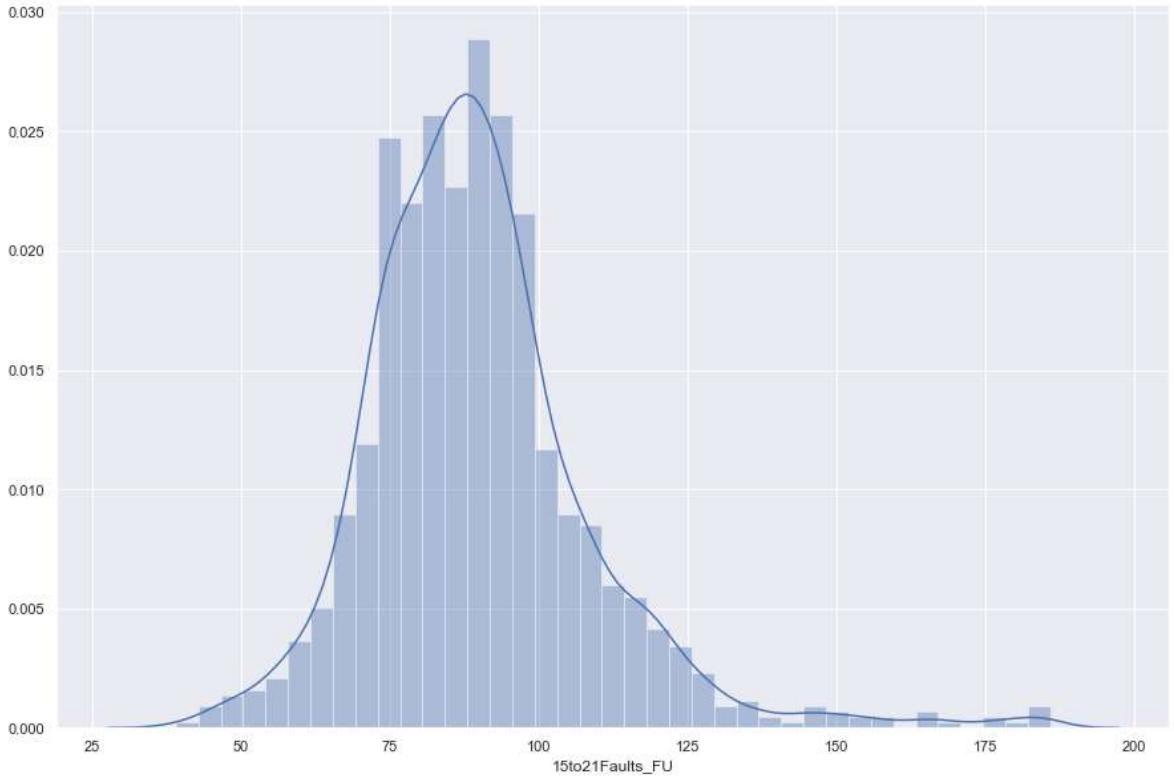


In [764]:

```
# Check Average Faults
plt.figure(figsize=(15,10))
plt.tight_layout()
sb.distplot(df_fu['15to21Faults_FU'])
```

Out[764]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x15c053c7b00>
```



In [765]:

```
# Find TRAIN Set Data for 15 to 21 Days for FU Faults
df_fu_train_np_rain=df_fu_train['Cummulative_Rain'].to_numpy()
df_fu_train_np_fault = df_fu_train['15to21Faults_FU'].to_numpy()

# Reshape test train data
df_fu_train_rain = df_fu_train_np_rain.reshape(len(df_fu_train_np_rain),1)
df_fu_train_fault = df_fu_train_np_fault.reshape(len(df_fu_train_np_fault),1)
```

In [766]:

```
lr = LinearRegression().fit(df_fu_train_rain,df_fu_train_fault)
```

In [767]:

```
lr.score(df_fu_train_rain, df_fu_train_fault)
```

Out[767]:

```
0.02427096447088306
```

In [768]:

```
# Find TEST Set Data for 15 to 21 Days for FU Faults
df_fu_test_np_rain=df_fu_test['Cummulative_Rain'].to_numpy()
df_fu_test_np_fault = df_fu_test['15to21Faults_FU'].to_numpy()

df_fu_test_rain = df_fu_test_np_rain.reshape(len(df_fu_test_np_rain),1)
df_fu_test_fault = df_fu_test_np_fault.reshape(len(df_fu_test_np_fault),1)
```

In [769]:

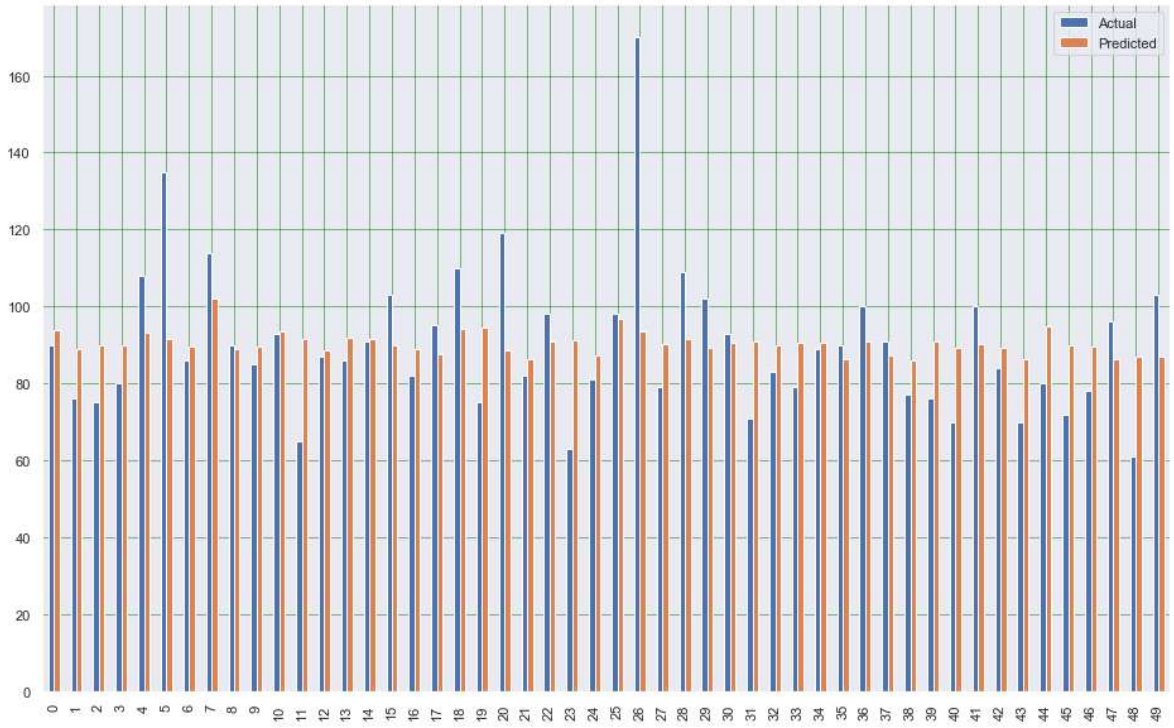
```
# testing model accuracy in test set
lr.score(df_fu_test_rain, df_fu_test_fault)
```

Out[769]:

```
0.02243310552540001
```

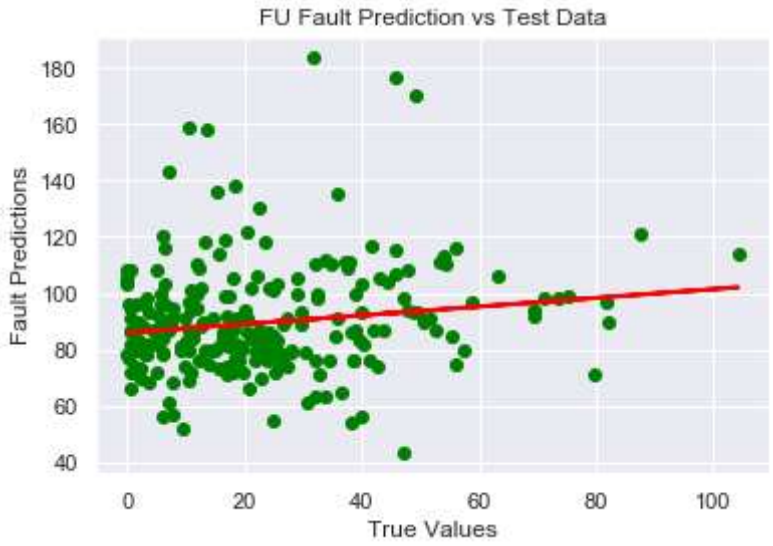
In [770]:

```
pred_fu_Fault_linear = lr.predict(df_fu_test_rain)
fu_actual_pred_linear= pd.DataFrame({'Actual': df_fu_test_fault.flatten(), 'Predicted'
bar_plot_df = fu_actual_pred_linear.head(50)
bar_plot_df.plot(kind='bar',figsize=(16,10))
plt.grid(which='major', linestyle='-', linewidth='0.5', color='green')
plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')
plt.show()
```



In [946]:

```
plt.scatter(df_fu_test_rain, df_fu_test_fault,  color='green')
plt.plot(df_fu_test_rain, pred_fu_Fault_linear, color='red', linewidth=2)
plt.title('FU Fault Prediction vs Test Data')
plt.xlabel('True Values')
plt.ylabel('Fault Predictions')
plt.show()
```



In [772]:

```
rmspe_fu = (np.sqrt(np.mean(np.square((df_fu_test_fault - pred_fu_Fault_linear) / df_f
print('Root Mean Squared Error Percentage for FU Faults: ', rmspe_fu,'%')
```

```
Root Mean Squared Error Percentage for FU Faults:  21.349197306932357 %
```

LN FAULT MODEL

Linear Regression

In [773]:

```
df_ln = df[['Cummulative_Rain','15to21Faults_LN']]
df_ln_train, df_ln_test = train_test_split(df_ln, test_size=0.2,random_state=1801)
```

In [955]:

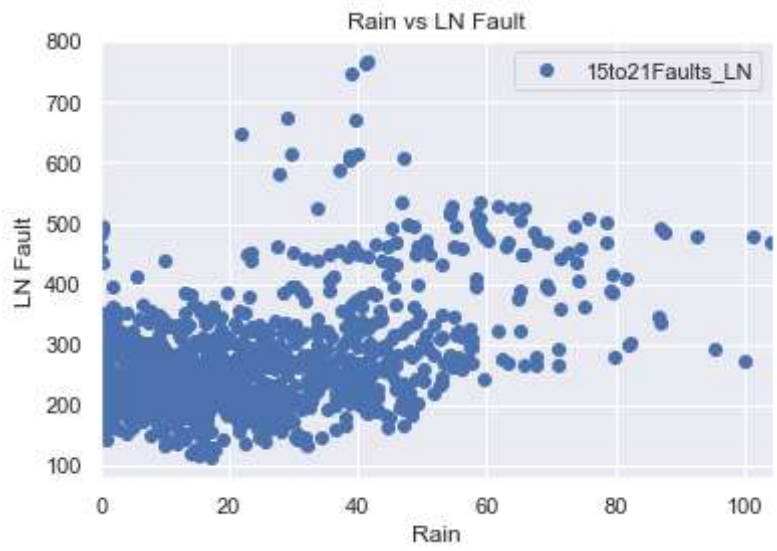
```
1 df_ln.corr()
```

Out[955]:

	Cummulative_Rain	15to21Faults_LN
Cummulative_Rain	1.000000	0.440418
15to21Faults_LN	0.440418	1.000000

In [774]:

```
1 df_ln.plot(x='Cummulative_Rain', y='15to21Faults_LN', style='o')
2 plt.title('Rain vs LN Fault')
3 plt.xlabel('Rain')
4 plt.ylabel('LN Fault')
5 plt.show()
```

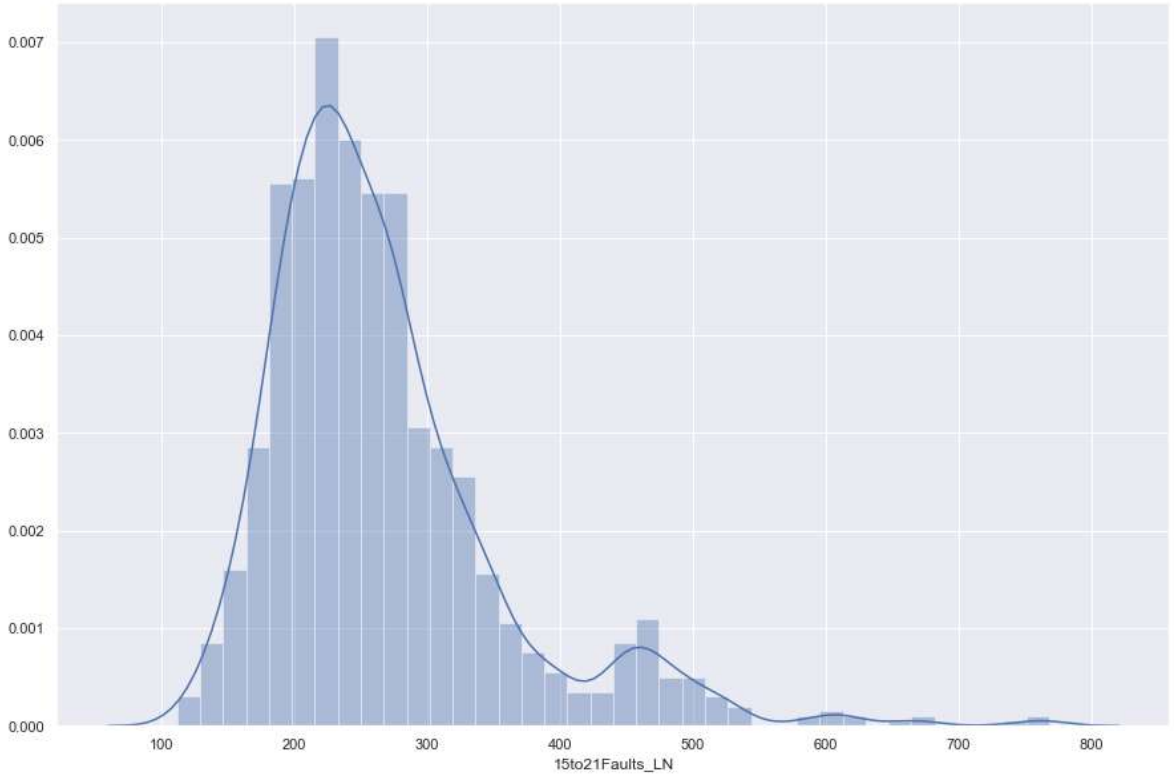


In [775]:

```
# Check Average Faults
plt.figure(figsize=(15,10))
plt.tight_layout()
sb.distplot(df_ln['15to21Faults_LN'])
```

Out[775]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x15c7d4c1908>
```



In [776]:

```
# Find TRAIN Set Data for 15 to 21 Days for LN Faults
df_ln_train_np_rain=df_ln_train['Cummulative_Rain'].to_numpy()
df_ln_train_np_fault = df_ln_train['15to21Faults_LN'].to_numpy()

# Reshape test train data
df_ln_train_rain = df_ln_train_np_rain.reshape(len(df_ln_train_np_rain),1)
df_ln_train_fault = df_ln_train_np_fault.reshape(len(df_ln_train_np_fault),1)
```

In [777]:

```
lr = LinearRegression().fit(df_ln_train_rain,df_ln_train_fault)
```

In [778]:

```
lr.score(df_ln_train_rain, df_ln_train_fault)
```

Out[778]:

```
0.18929831487567994
```

In [779]:

```
# Find TEST Set Data for 15 to 21 Days for FU Faults
df_ln_test_np_rain=df_ln_test['Cummulative_Rain'].to_numpy()
df_ln_test_np_fault = df_ln_test['15to21Faults_LN'].to_numpy()

df_ln_test_rain = df_ln_test_np_rain.reshape(len(df_ln_test_np_rain),1)
df_ln_test_fault = df_ln_test_np_fault.reshape(len(df_ln_test_np_fault),1)
```

In [780]:

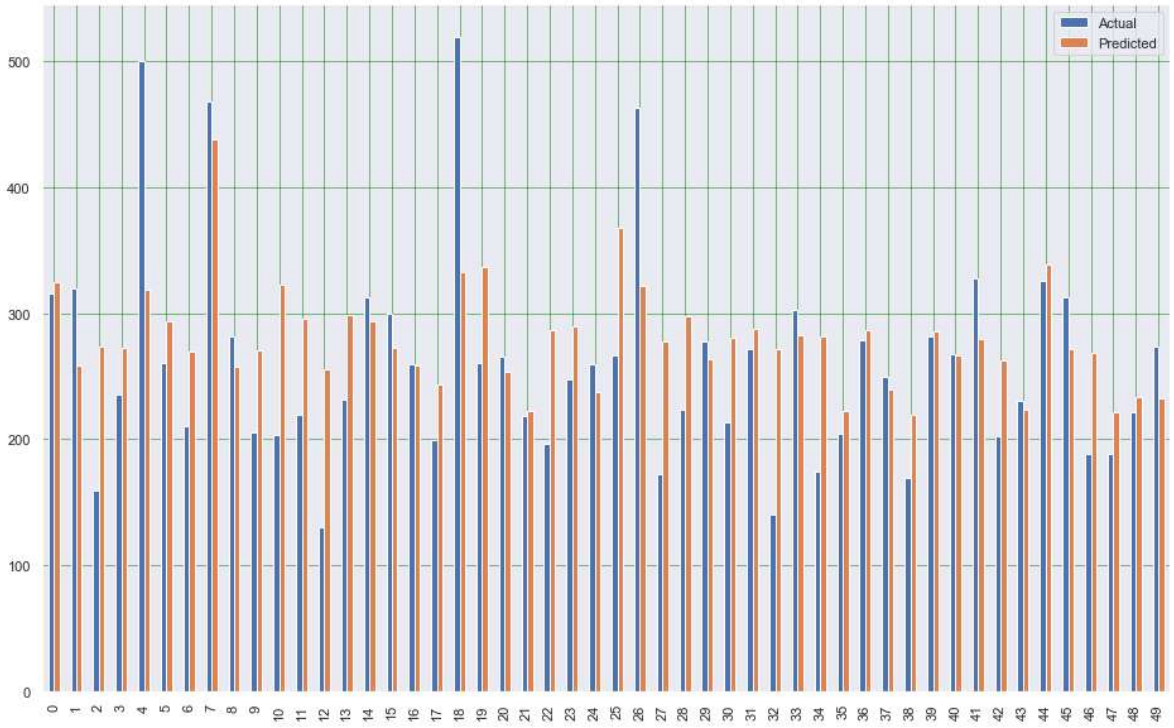
```
# testing model accuracy in test set
lr.score(df_ln_test_rain, df_ln_test_fault)
```

Out[780]:

```
0.21384722987389126
```

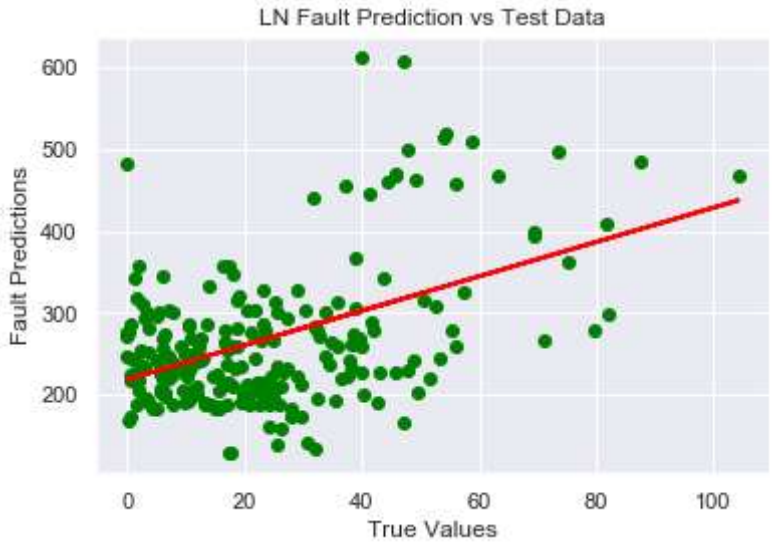
In [781]:

```
pred_ln_Fault_linear = lr.predict(df_ln_test_rain)
ln_actual_pred_linear= pd.DataFrame({'Actual': df_ln_test_fault.flatten(), 'Predicted'
bar_plot_df = ln_actual_pred_linear.head(50)
bar_plot_df.plot(kind='bar',figsize=(16,10))
plt.grid(which='major', linestyle='-', linewidth='0.5', color='green')
plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')
plt.show()
```



In [947]:

```
plt.scatter(df_ln_test_rain, df_ln_test_fault,  color='green')
plt.plot(df_ln_test_rain, pred_ln_Fault_linear, color='red', linewidth=2)
plt.title('LN Fault Prediction vs Test Data')
plt.xlabel('True Values')
plt.ylabel('Fault Predictions')
plt.show()
```



In [783]:

```
rmspe_ln = (np.sqrt(np.mean(np.square((df_ln_test_fault - pred_ln_Fault_linear) / df_l
print('Root Mean Squared Error Percentage for LN Faults: ', rmspe_ln,'%')
```

```
Root Mean Squared Error Percentage for LN Faults:  29.48329493509179 %
```

OTHER FAULT MODEL

Linear Regression

In [784]:

```
df_other = df[['Cummulative_Rain','15to21Faults_OTHER']]
df_other_train, df_other_test = train_test_split(df_other, test_size=0.2,random_state=
```

In [956]:

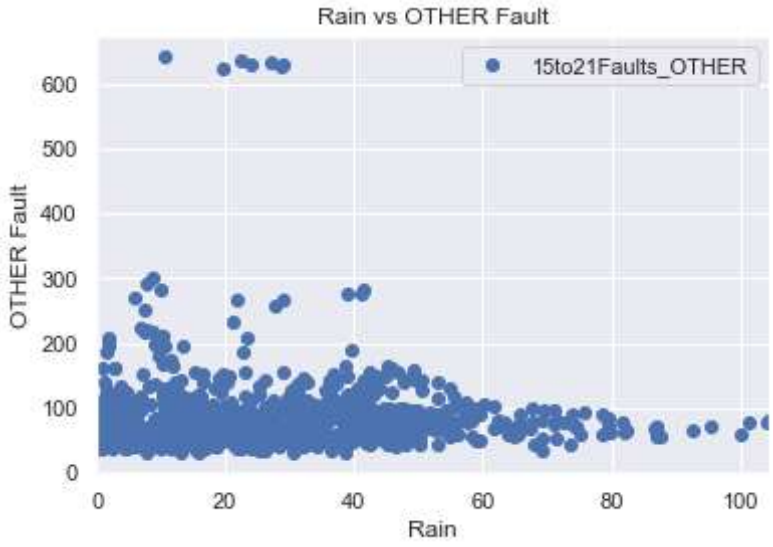
```
df_other.corr()
```

Out[956]:

	Cummulative_Rain	15to21Faults_OTHER
Cummulative_Rain	1.000000	0.035445
15to21Faults_OTHER	0.035445	1.000000

In [785]:

```
df_other.plot(x='Cummulative_Rain', y='15to21Faults_OTHER', style='o')
plt.title('Rain vs OTHER Fault')
plt.xlabel('Rain')
plt.ylabel('OTHER Fault')
plt.show()
```

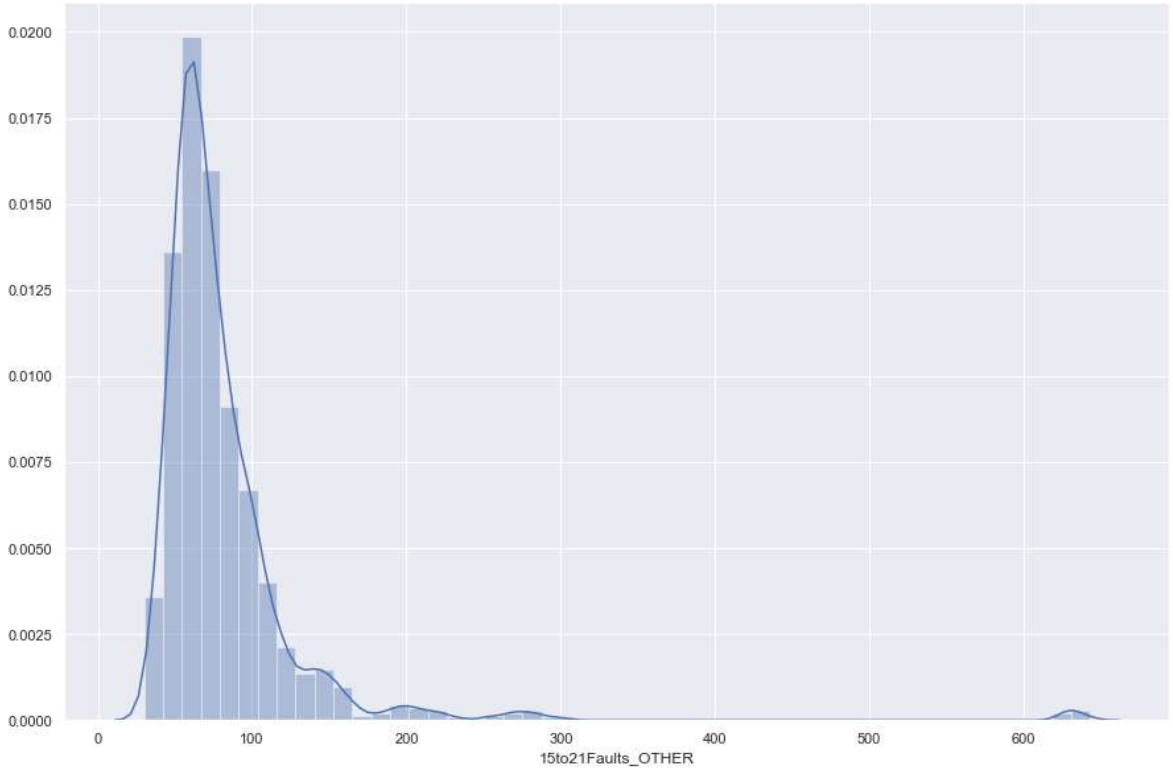


In [786]:

```
# Check Average Faults
plt.figure(figsize=(15,10))
plt.tight_layout()
sb.distplot(df_other['15to21Faults_OTHER'])
```

Out[786]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x15c7d35bf28>
```



In [787]:

```
# Find TRAIN Set Data for 15 to 21 Days for OTHER Faults
df_other_train_np_rain=df_other_train['Cummulative_Rain'].to_numpy()
df_other_train_np_fault = df_other_train['15to21Faults_OTHER'].to_numpy()

# Reshape test train data
df_other_train_rain = df_other_train_np_rain.reshape(len(df_other_train_np_rain),1)
df_other_train_fault = df_other_train_np_fault.reshape(len(df_other_train_np_fault),1)
```

In [788]:

```
lr = LinearRegression().fit(df_other_train_rain,df_other_train_fault)
```

In [789]:

```
lr.score(df_other_train_rain, df_other_train_fault)
```

Out[789]:

```
0.0014555832446628836
```

In [790]:

```
# Find TEST Set Data for 15 to 21 Days for OTHER Faults
df_other_test_np_rain=df_other_test['Cummulative_Rain'].to_numpy()
df_other_test_np_fault = df_other_test['15to21Faults_OTHER'].to_numpy()

df_other_test_rain = df_other_test_np_rain.reshape(len(df_other_test_np_rain),1)
df_other_test_fault = df_other_test_np_fault.reshape(len(df_other_test_np_fault),1)
```

In [791]:

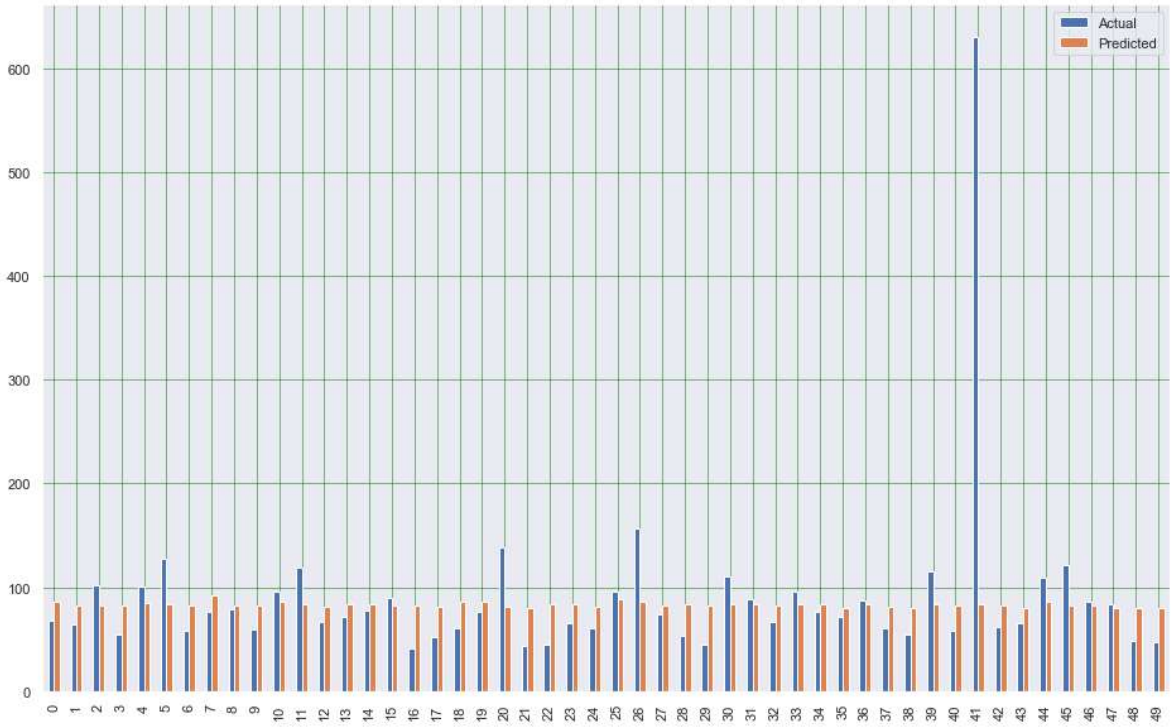
```
# testing model accuracy in test set
lr.score(df_other_test_rain, df_other_test_fault)
```

Out[791]:

```
-0.007530650746826684
```

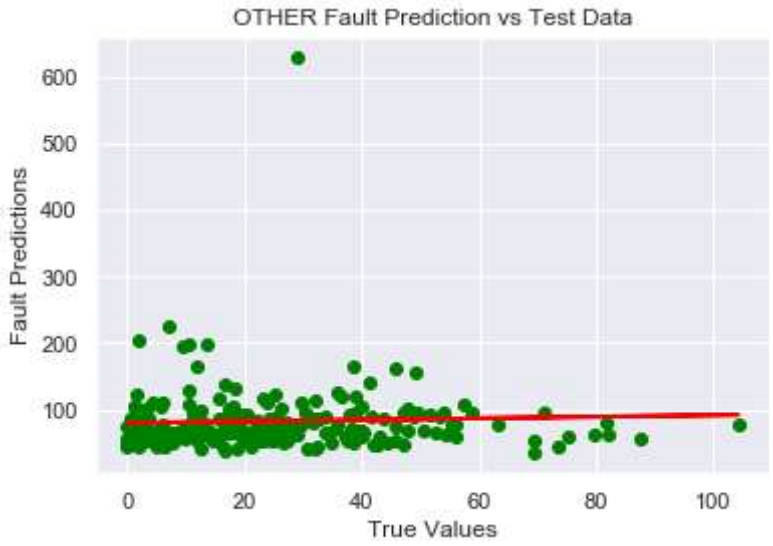
In [792]:

```
pred_other_Fault_linear = lr.predict(df_other_test_rain)
other_actual_pred_linear= pd.DataFrame({'Actual': df_other_test_fault.flatten(), 'Pred
bar_plot_df = other_actual_pred_linear.head(50)
bar_plot_df.plot(kind='bar',figsize=(16,10))
plt.grid(which='major', linestyle='-', linewidth='0.5', color='green')
plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')
plt.show()
```



In [948]:

```
plt.scatter(df_other_test_rain, df_other_test_fault,  color='green')
plt.plot(df_other_test_rain, pred_other_Fault_linear, color='red', linewidth=2)
plt.title('OTHER Fault Prediction vs Test Data')
plt.xlabel('True Values')
plt.ylabel('Fault Predictions')
plt.show()
```



In [794]:

```
rmspe_other = (np.sqrt(np.mean(np.square((df_other_test_fault - pred_other_Fault_linea
print('Root Mean Squared Error Percentage for Other Faults: ', rmspe_other,'%')
```

```
Root Mean Squared Error Percentage for Other Faults:  43.38603966237854 %
```