In Class Part

Q1: Which statement is TRUE about Margin of Error ?

A) Its value is the same for the same confidence levels
B) For smaller significance levels, there will be smaller margins of error
C) It directly depends on standard deviation
D) It is independent of the type of distribution

In [2]:
```
#Answer is C
```

Q2: What is the Z score associated with 90% confidence interval?

A) 1.645
B) 1.960
C) 2.576
D) 2.807

In [3]:
```
#Answer is A
```

Q3:In A/B Testing of a new web design, the analysts are randomly sending some users to the original website and some users to the newly designed version. Why not just change the website and monitor it for a month?

A) To avoid extra costs
B) To control all other variables
C) To expedite the analysis process
D) None of the above

In [4]:
```
#Asnwer is B
```

Q4: MSE (Mean Squared Error) is one of the ways to estimate the parameters of a linear regression model.

A) TRUE
B) FALSE

In [5]:
```
#Asnwer is B | FALSE
```

Q5: What is correct about the general format of a linear regression model: $Ye = \beta X + \alpha$

A) Y_e is the predicted values of Y based on the estimated parameters.
B) α is the slope the linear regression model
C) β is the intercept of the linear regression model
D) All of the above

In [6]:
```
#Asnwer is A
```

Q6: Which one is NOT correct about covariance?

A) It is a measure of the joint variability of two random variables
B) Its absolute value shows the strength of relationship between two variables
C) It shows the direction of relationship between two variables
D) None of the above

In [7]:
```
# Answer is B
```

Q7: Pearson's r ...

A) Is a normalized version of variance
B) Is a value between 0 and 1
C) Is the most popular type of correlation coefficients
D) All of the above

In [8]:
```
# Answer is C
```

Q8: Using a multiple linear regression model, multiple targets can be predicted using the same set of predictors.

A) TRUE
B) FALSE

In [9]:
```
# Answer is B | False
```

Q9: Which one is NOT a goal of linear regression modeling?

A) To predict a target variable using predictor variables
B) To explore the relationship between different predictors and the output variable
C) To check whether a set of samples follow a normal distrubtion
D) None of the above

In [10]:
```
# Answer is C
```

Q10: We use Goodness-of-Fit metrics to assess the performance of regression and classification models.

A) TRUE
B) FALSE

In [11]:
```
# Answer is A | TRUE
```

Q11: Which one is different?

A) RMSE
B) r
C) R2
D) MAE

In [12]:
```
# Answer is B
```

Q12: Which one cannot be infered by visual assessment of linear regression models?

A) The quality of fit
B) The range(s) where the model is more reliable
C) The slope and coefficient of the linear fit
D) The bias-variance tradeoff

In [13]:
```
# Answer is D
```

Q13: Which statement is TRUE?

A) Low bias and low variance can be achieved if the model is nonlinear
B) A very simplisitic model results is high bias
C) When a small change in predictors cause a large change in target, the bias is high
D) A good model has low bias and high variance

In [14]:
```
# Answer is B
```

Q14: Which one is not a component of Kling-Gupta Efficiency?

A) The Nash-Sutcliffe efficiency coefficient
B) The Pearson product-moment correlation coefficient
C) The ratio between the mean of the simulated values and the mean of the observed ones.
D) B and C

In [15]:
```
# Answer is A
```

Q15: The confidence band of a linear regression model tells us where the true linear regression line of the population will lie within the confidence interval of the regression line calculated from the sample data.

A) TRUE
B) FALSE

In [16]:
```
# Answer is A | TRUE
```

Q16: For any specific value, the confidence interval is more meaningful than the prediction interval.

A) TRUE
B) FALSE

In [17]:
```
# Answer is B | FALSE
```

Q17: Which statement is TRUE about classification vs. regression?

A) If the predictors are discrete, it is a classification problem.
B) If one of the predictors or the target is discrete, it is a classification problem.
C) If the target variable is discrete, it is a classification problem.
D) If the target is numeric, it is a regression problem.

In [18]:
```
# Answer is C
```

Q18: Which one is NOT true about logistic regression?

A) Is an algorithm for both regression and classification problems
B) Requires a cutoff value to decide which class each case belongs to
C) Can be applied to binary problems as well as multiclass classifications.
D) Maximum Likelihood Estimation is used to estimate the best values for its coefficients.

In [19]:
```
# Answer is A
```

Q19: Which one is TRUE about the logistic function?

A) It is also known as the likelihood function
B) It looks like an inverted "S"
C) It maps any real valued number onto a value between 0 and 1
D) Most of its values are around 0.5

In [20]:
```
# Answer is C
```

Q20: Which one is NOT true about discrete GOF metrics?

A) The accuracy is the ratio of the number of correct predictions to the total number of predictions (or observations)
B) F1-score is the arithmetic mean of the Precision and Recall
C) The sensitivity is the ratio of the number of true positives to the number of actual positives.
D) The specificity is the ratio of the number of true negatives to the number of actual negatives

In [21]:
```
# Answer if B
```

In [0]:

In [0]:

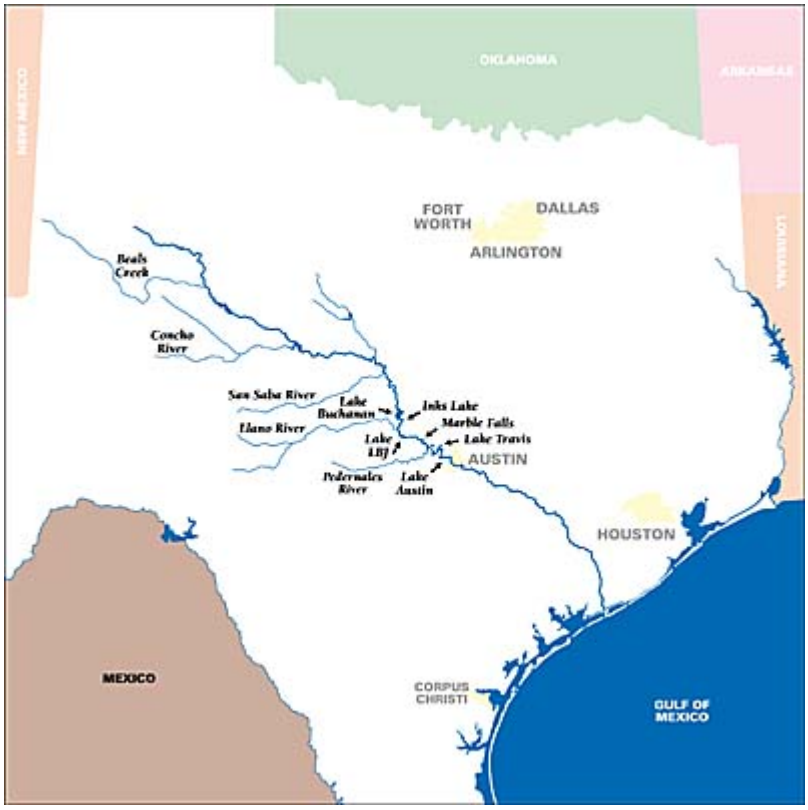
In [ ]:

In [ ]:

Take-Home Portion

Exercise1: Streamflow Forecasting in Colorado River, TX.

The Colorado River is the 18th longest river in the United States and the longest river wholly located in Texas. The Colorado River originates from the Llano Estacado region in west Texas with an elevation of 1000 meters, flows southeast through eleven major reservoirs (e.g. Lake J.B. Thomas, E.V. Spence Reservoir and Lake O.H. Ivie) and eventually empties into the Gulf of Mexico near Matagorda. The river is considered as the life blood of Texas due to its essential role for the state's economy, environment, agriculture, power production and developing municipalities and industries. The Colorado River begin its journey south of Lubbock as an intermittent stream. This means that there are periods where the riverbed goes dry and the flowrate is lower than measurable amounts.



The "Colorado River Data.csv" dataset contains monthly streamflow recordings from 03/1988 to 03/2019 at USGS station #08117995 near Gail, Borden County, Texas as well as several hydro-meteorological variables from PRISM Climate Data:

Columns	Info.
Date	Date of recording in YYYY-MM format
PPT	Accumulated Monthly Precipitation (mm)
Tmin	Minimum Recorded Temperature (degrees C)
Tav	Average Recorded Temperature (degrees C)
Tmax	Maximum Recorded Temperature (degrees C)
delt	Difference between Maximum and Minimum Recorded Temperatures (degrees C)
SMI	Soil Moisture Index (SMI) - The product of precipitation and average temperature
ET	Evapotranspiration (mm)
Flow_cfs	Average recorded flowrate (cfs)
Flow?	Flow status: It is 1 when there is a flow and it is 0 when there is no flow.

Follow the steps and answer the following questions:

- Step1: Read the "Colorado River Data.csv" file as a dataframe. Explore the dataframe and in a markdown cell breifly describe the different variables in your own words.
- Step2: Get the following plots and analyze each one of them:
  - Plot a histogram with KDE for precipitation. | What does it show? What are the most common value you expect to see based on this graph?

    - Plot flow vs. date. | What are the most noteable things you see in this graph? What are some extreme values and why do you think they have happened?
    - Plot a joint KDE plot with ET on x-axis and flow status on the y-axis. |What does this graph show you? Do you think evaporation can be a good predictor for flow status?
- Step3: Calculate and compare the correlation coefficient of the variables. Analyze the results and decide.
    - which parameters have the strongest relationship and weakest relationship with the flowrates? Do you identify correlation and causation or just correlation?
    - what predictor has the strongest relationship with ET?
    - which parameter has a negative correlation with precipitation?
- Step4: Think about a few enginreeing applications in which we may need to forecast the flowrate in a river?
- Step5: Use linear regression modeling in primitive python, get the linear model's coefficients via Ordinary Least Squares methods, make a plot and VISUALLY assess the quality of a linear fit with precipitation as the predictor, and flowrate as outcome. Then, use RMSE, Pearson's r, and R2 to describe the performance of your model. Explain the results of this analysis in a markdown cell.
- Step6: Use multiple linear regression modeling with scikit-learn and use all "ET","SMI", and "delt" to predict the flowrates. Then, use RMSE, Pearson's r, and R2 to describe the performance of your model. Explain the results of this analysis in a markdown cell.
- Step7: Think about a few enginreeing applications in which we may need to forecast the flow status (flow/no flow) in a river?
- Step8: Use logistic regression and "PPT","SMI", and "delt" as predictors to predict the status of flow. Use a 75/25 split for training and testing. Then, get the confusion matrix and use classification_report to describe the performance of your model. Also, get a heatmap and visually assess the predictions of your model. Calculate accuracy, recall, precision, and F1-score for your model. Explain the results of this analysis in a markdown cell.

- Step9: Was this a balanced classification problem? why?

In [51]:
```
#Step0: Load the necessary libraries
import numpy as np
import pandas as pd
import statistics
import scipy.stats
import matplotlib.pyplot
from matplotlib import pyplot as plt
import seaborn as sns

import statsmodels.formula.api as smf
import sklearn.metrics as metrics
```

In [35]:
```
#Step1:Read the "Colorado River Data.csv" file as a dataframe
df = pd.read_csv('Colorado River Data.csv')
df.head()
```

Out[35]:

	Date	PPT	Tmin	Tav	Tmax	delt	SMI	ET	Flow_cfs	Flow?
0	1988-03	16.09	2.8	12.3	21.8	19.0	197.907	31.429845	2.68	1
1	1988-04	17.63	7.9	17.0	26.1	18.2	299.710	61.427057	2.33	1
2	1988-05	63.44	14.1	21.4	28.7	14.6	1357.616	104.931753	4.56	1
3	1988-06	58.01	18.6	25.8	33.0	14.4	1496.658	149.374517	18.20	1
4	1988-07	91.98	20.5	26.6	32.7	12.2	2446.668	161.751506	76.10	1

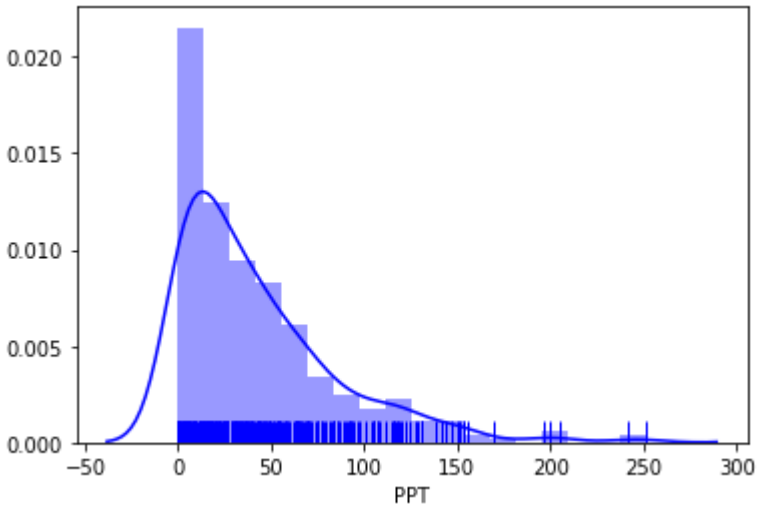
In [36]:
```
# Explore the dataframe: Describe the df
df.describe()
```

Out[36]:

	PPT	Tmin	Tav	Tmax	delt	SMI	ET	F
count	373.000000	373.000000	373.000000	373.000000	373.000000	373.000000	373.000000	373
mean	42.483217	10.456568	17.771046	25.087668	14.631099	862.258668	80.490330	13
std	42.894824	8.182669	7.896588	7.715736	1.889885	1014.255443	63.165859	51
min	0.000000	-4.600000	3.300000	8.800000	9.900000	0.000000	2.151330	0
25%	9.910000	2.500000	10.500000	18.200000	13.400000	126.100000	19.647938	0
50%	30.700000	10.400000	18.100000	25.900000	14.500000	428.280000	65.127493	0
75%	62.850000	18.400000	25.300000	31.800000	16.000000	1389.024000	138.843235	7
max	251.240000	24.300000	31.500000	38.800000	19.400000	5672.665000	218.125353	709

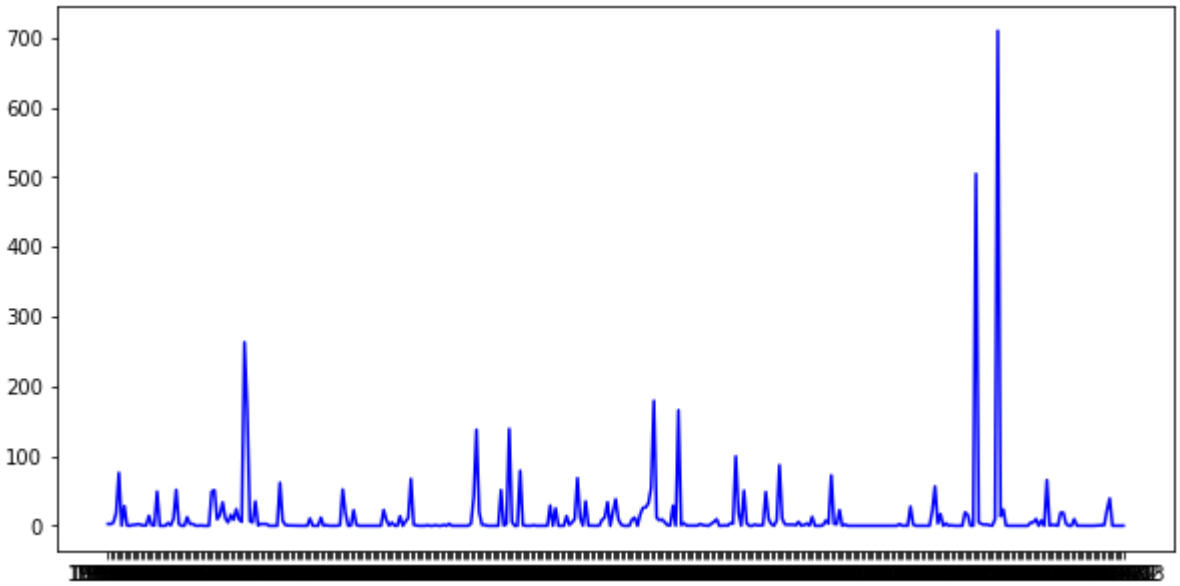
In [41]:
```
#Step2-part A:
sns.distplot(df['PPT'], kde = True, rug= True, color ='blue')
```

Out[41]: <matplotlib.axes._subplots.AxesSubplot at 0x1d435ae7dc8>



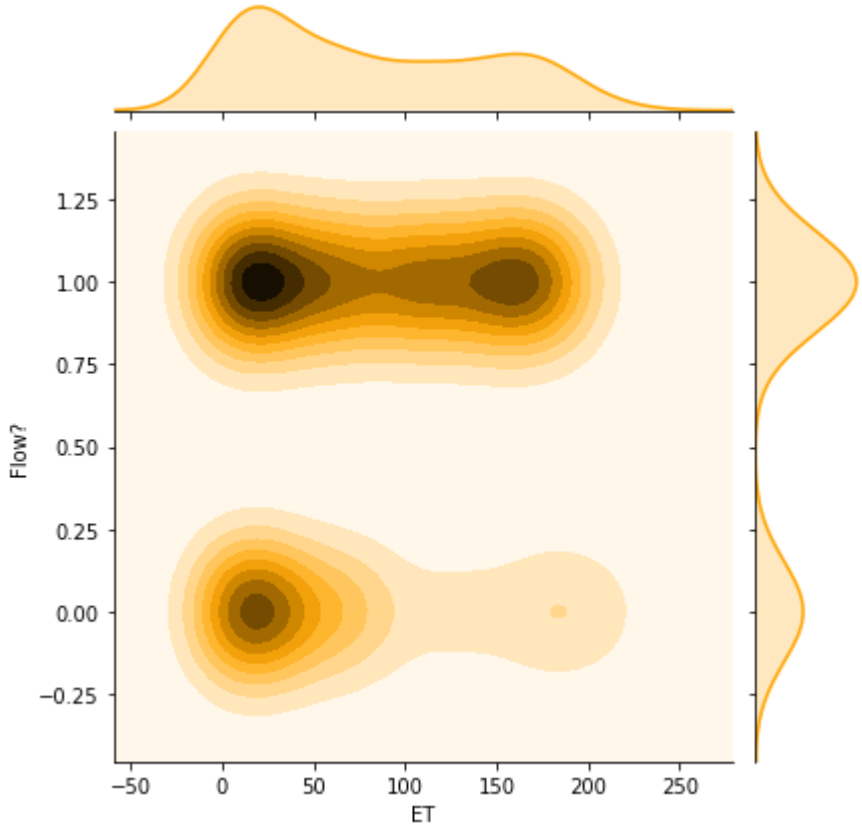
In [52]:
```
#Step2-part B:
date = df['Date']
flow = df['Flow_cfs']

# Plot here
myfigure = matplotlib.pyplot.figure(figsize = (10,5)) # generate an object fro
m the figure class, set aspect ratio

matplotlib.pyplot.plot(date,flow, color ='blue')
```

Out[52]: [<matplotlib.lines.Line2D at 0x1d43a0add88>]



In [54]:
```
#Step2-part C:
sns.jointplot(x ='ET', y ='Flow?', data = df, kind ='kde',color="orange")
```

Out[54]:
```
<seaborn.axisgrid.JointGrid at 0x1d439da2308>
```



In [26]:
```
#Step3: Calculate and compare the correlation coefficient
#What can we infer?
df.corr(method ='pearson')
```

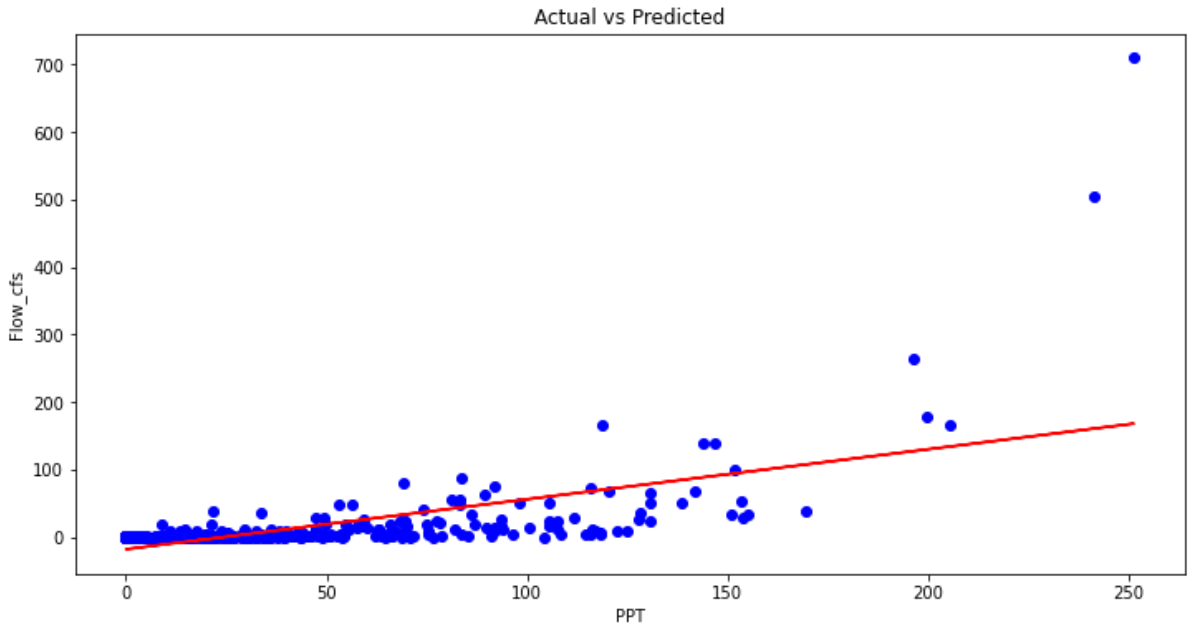
Out[26]:

	PPT	Tmin	Tav	Tmax	delt	SMI	ET	Flow_cfs
PPT	1.000000	0.377617	0.317593	0.249061	-0.618147	0.938977	0.289487	0.614588
Tmin	0.377617	1.000000	0.993709	0.973441	-0.355501	0.572704	0.974184	0.151265
Tav	0.317593	0.993709	1.000000	0.992938	-0.248663	0.526083	0.975157	0.123433
Tmax	0.249061	0.973441	0.992938	1.000000	-0.132076	0.468836	0.962877	0.092345
delt	-0.618147	-0.355501	-0.248663	-0.132076	1.000000	-0.565551	-0.286853	-0.277924
SMI	0.938977	0.572704	0.526083	0.468836	-0.565551	1.000000	0.502235	0.575503
ET	0.289487	0.974184	0.975157	0.962877	-0.286853	0.502235	1.000000	0.101266
Flow_cfs	0.614588	0.151265	0.123433	0.092345	-0.277924	0.575503	0.101266	1.000000
Flow?	0.452067	0.234497	0.197389	0.155013	-0.382443	0.423881	0.184756	0.175355

In [ ]:
```
#Step4:
```

In [58]:
```
#Step5:
# Calculate the mean of X and y
xmean = np.mean(df['PPT'])
ymean = np.mean(df['Flow_cfs'])

# Calculate the terms needed for the numator and denominator of beta
df['xycov'] = (df['PPT'] - xmean) * (df['Flow_cfs'] - ymean)
df['xvar'] = (df['PPT'] - xmean)**2

# Calculate beta and alpha
beta = df['xycov'].sum() / df['xvar'].sum()
alpha = ymean - (beta * xmean)
print(f'alpha = {alpha}')
print(f'beta = {beta}')

X = np.array(df['PPT'])
Y = np.array(df['Flow_cfs'])

ypred = alpha + beta * X
# Plot regression against actual data
plt.figure(figsize=(12, 6))
plt.plot(X, Y, 'ro', color="blue")   # scatter plot showing actual data
plt.plot(X, ypred, color="red")     # regression line
plt.title('Actual vs Predicted')
plt.xlabel('PPT')
plt.ylabel('Flow_cfs')

plt.show()
###
#GOF metrics:
print("RMSE for PPT as predictor is ",np.sqrt(metrics.mean_squared_error(Y, yp
red)))
print("R2 for PPT as predictor is ",metrics.r2_score(Y, ypred))

pearson_r = pearsonr(ypred, Y)

print("Pearson's r for PPT as predictor is ",pearson_r[0])
```

```
alpha = -18.333691604121523
beta = 0.7411679241433508
```




```
RMSE for PPT as predictor is  40.751870999819324
R2 for PPT as predictor is  0.37771885033965624
Pearson's r for PPT as predictor is  0.6145883584478771
```

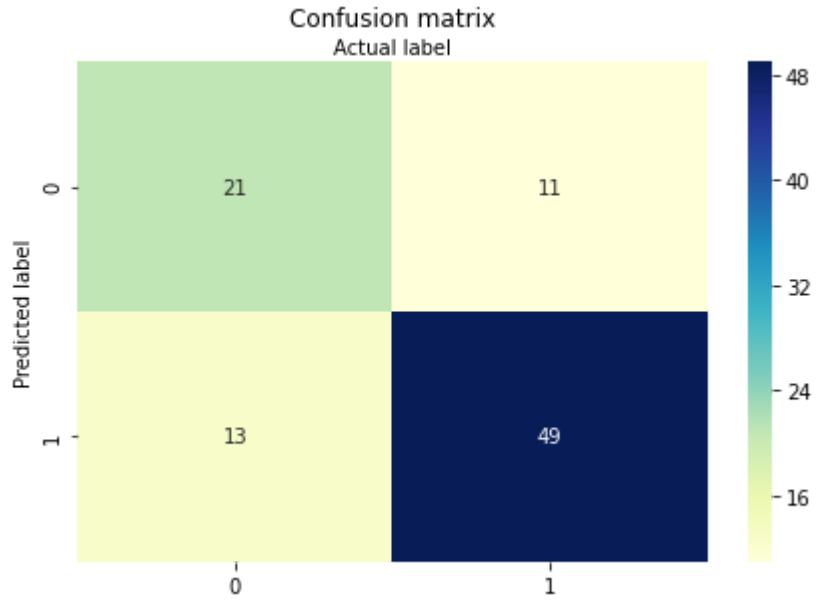
In [62]:
```
#Step6:
from sklearn.linear_model import LinearRegression

# Split data into predictors X and output Y
predictors = ['ET', 'SMI', 'delt']
X = df[predictors]
Y = df['Flow_cfs']

# Initialise and fit model
lm = LinearRegression()
model = lm.fit(X, Y)

# Predict values
big_pred = model.predict(X)
#GOF metrics:
print("RMSE for PPT as predictor is ",np.sqrt(metrics.mean_squared_error(Y, bi
g_pred)))
print("R2 for PPT as predictor is ",metrics.r2_score(Y, big_pred))

pearson_r = pearsonr(big_pred, Y)

print("Pearson's r for PPT as predictor is ",pearson_r[0])
```

```
RMSE for PPT as predictor is  40.62517214384142
R2 for PPT as predictor is  0.38158221887758004
Pearson's r for PPT as predictor is  0.6177234161642086
```

In [ ]:
```
#Step7:
```

In [68]:
```
#Step8:
#split dataset in features and target variable
feature_cols = ['PPT', 'SMI', 'delt']
X = df[feature_cols] # Features
Y = df['Flow?'] # Target variable
# split X and y into training and testing sets
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(X,Y,test_size=0.25,random_state
=0)
# import the class
from sklearn.linear_model import LogisticRegression

# instantiate the model (using the default parameters)
#logreg = LogisticRegression()
logreg = LogisticRegression()
# fit the model with data
logreg.fit(X_train,y_train)

#
y_pred=logreg.predict(X_test)

# import the metrics class
from sklearn import metrics
cnf_matrix = metrics.confusion_matrix(y_pred, y_test)
print(cnf_matrix)
# Visualize
class_names=[0,1] # name  of classes
fig, ax = plt.subplots()
tick_marks = np.arange(len(class_names))
plt.xticks(tick_marks, class_names)
plt.yticks(tick_marks, class_names)
# create heatmap
sns.heatmap(pd.DataFrame(cnf_matrix), annot=True, cmap="YlGnBu" ,fmt='g')
ax.xaxis.set_label_position("top")
plt.tight_layout()
plt.title('Confusion matrix', y=1.1)
plt.ylabel('Predicted label')
plt.xlabel('Actual label')
#
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
print("Precision:",metrics.precision_score(y_test, y_pred))
print("Recall:",metrics.recall_score(y_test, y_pred))
print("F1-score:",metrics.f1_score(y_test, y_pred))
#
from sklearn.metrics import classification_report

print(classification_report(y_test, y_pred))
```

```
[[21 11]
 [13 49]]
Accuracy: 0.7446808510638298
Precision: 0.7903225806451613
Recall: 0.8166666666666667
F1-score: 0.8032786885245902
              precision    recall  f1-score   support

           0       0.66      0.62      0.64        34
           1       0.79      0.82      0.80        60

    accuracy                           0.74        94
   macro avg       0.72      0.72      0.72        94
weighted avg       0.74      0.74      0.74        94


C:\Users\Farha\Anaconda3\lib\site-packages\sklearn\linear_model\logistic.py:432: FutureWarning: Default solver will be changed to 'lbfgs' in 0.22. Specify a solver to silence this warning.
  FutureWarning)
```



In [ ]:
```
#Step9:
```

In [ ]:

In [ ]:

Take-Home Part | Alternatives for Bonus Questions

In [0]: