Project: Streamflow Analysis

Why Streamflow Analysis is Important?

Information gained from streamflow data is used for many different purposes:

- ***Water supply plans***
  - Municipalities
  - Agriculture
  - Industries
- ***Engineering design***
  - Reservoirs
  - Bridges, Roads, Culverts
  - Treatment Plans
- ***Operations***
  - Reservoirs
  - Power plants
  - Navigation
- ***Identifying changes in streamflows***
  - Climate change
  - Water use
  - Land use
- ***Flood planning and warning***
  - Floodplain mapping
  - Flood forecasts
- ***Streamflow forecasting***
- ***Characterizing and evaluating in-stream conditions***
  - Habitat assessment
  - Environmental flow requirements
  - Recreation
- ***Support of water quality sampling***
  - Water quality conditions
  - Contaminant transport

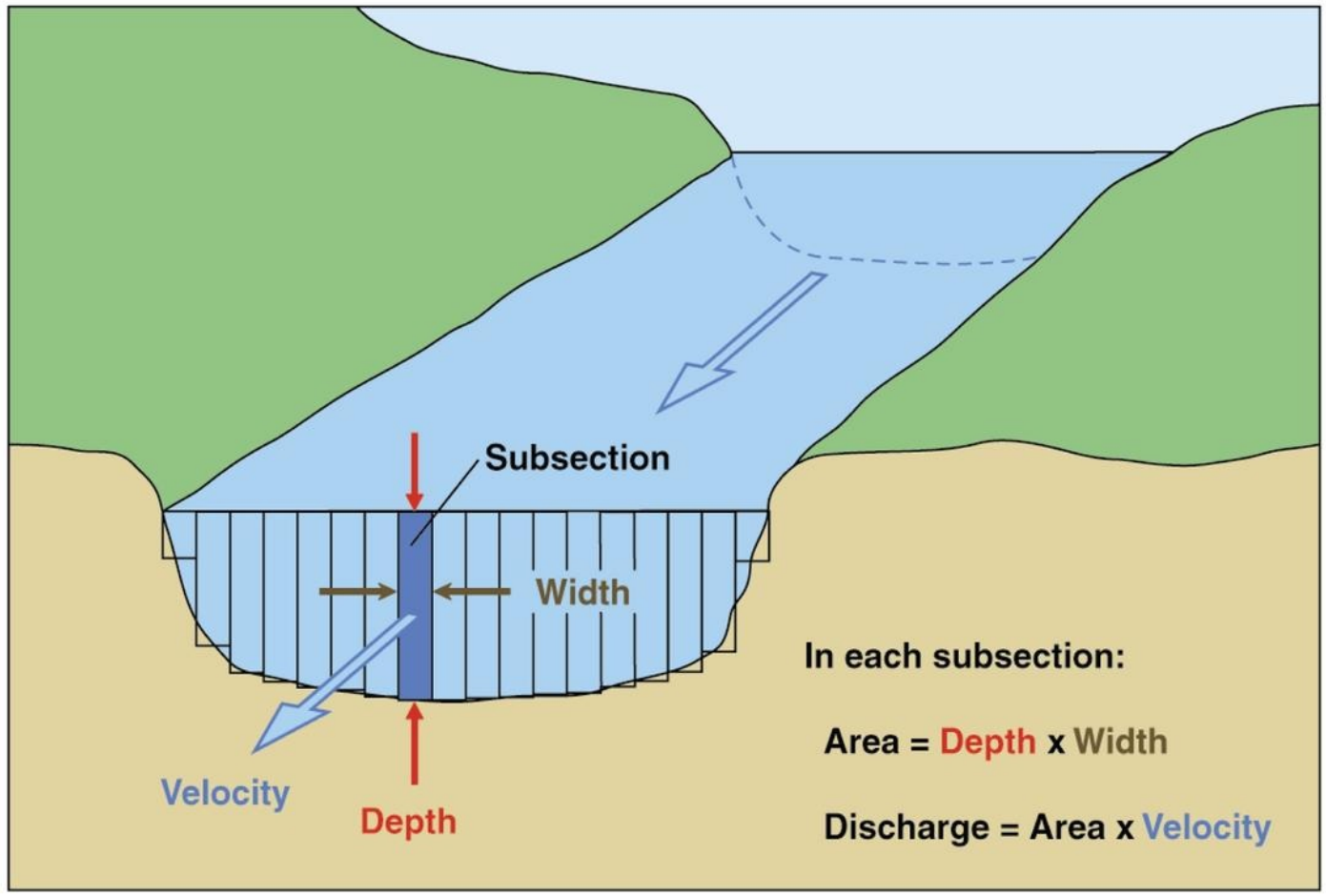
In this project, you will analyse a streamflow dataset and build a Machine Learning Model to predict the flow status and flowrate in a river.

How Streamflow is Measured?

USGS has an interesting narration on the process that you can find here (https://www.usgs.gov/special-topic/water-science-school/science/how-streamflow-measured?qt-science_center_objects=0#qt-science_center_objects)

Streamgaging generally involves 3 steps:

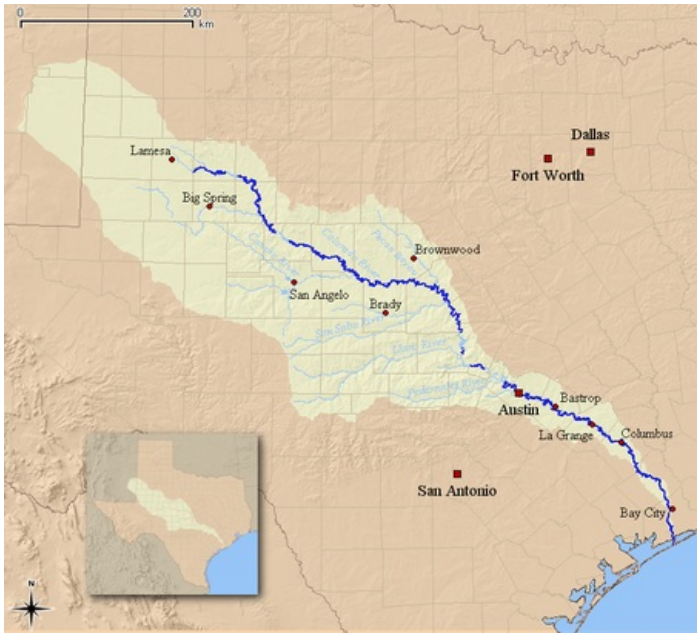
1. ***Measuring stream stage***—obtaining a continuous record of stage—the height of the water surface at a location along a stream or river
2. ***The discharge measurement***—obtaining periodic measurements of discharge (the quantity of water passing a location along a stream)
3. ***The stage-discharge relation***—defining the natural but often changing relation between the stage and discharge; using the stage-discharge relation to convert the continuously measured stage into estimates of streamflow or discharge



The Case Study: The Colorado River in Texas

The Colorado River is an approximately 862-mile (1,387 km) long river in the U.S. state of Texas. It is the 18th longest river in the United States and the longest river with both its source and its mouth within Texas.

The Colorado River originates south of Lubbock, on the Llano Estacado near Lamesa. It flows generally southeast out of the Llano Estacado and through the Texas Hill Country, then through several reservoirs including Lake J.B. Thomas, E.V. Spence Reservoir, and O.H. Ivie Lake. The river flows through several more reservoirs before reaching Austin, including Lake Buchanan, Inks Lake, Lake Lyndon B. Johnson (commonly referred to as Lake LBJ), and Lake Travis. The Llano River joins the Colorado at Lake LBJ near Kingsland, and the Pedernales River joins at Lake Travis near Briarcliff. After passing through Austin, the Colorado River continues flowing southeast until emptying into Matagorda Bay on the Gulf of Mexico, near Matagorda. The Colorado is the largest river lying entirely within Texas; it drains an area of about 39,900 square miles (103,350 square km) and receives several forks of the Concho River, the Pecan Bayou, and the San Saba, Llano, and Pedernales rivers.



The river is an important source of water for farming, cities, and electrical power production. In addition to power plants operating on each of the major lakes, waters of the Colorado are used for cooling the South Texas Nuclear Project near Bay City. Altogether, there are over 7,500 miles of creeks, streams, and rivers in our basin, and well over 2 million people live and work here. The Colorado's watershed includes several major metropolitan areas, including Midland-Odessa, San Angelo, and Austin, and there are hundreds of smaller towns and communities as well. Many communities, like Austin, rely on the Colorado River for 100% of their municipal water. Because of its importance to the state's economy, environment, industry, and agriculture it is recognized as the lifeblood of Texas.

/cdn.vox-cdn.com/uploads/chorus_image/image/64045588/Mount_Bonnell_shutterstock.0.jpg)

The Dataset | Streamflow Data

Streamflow data is downloaded for the most upstream streamflow monitoring station from "waterdata.usgs.gov". USGS Streamflow Monitoring Station 08117995 located in Borden County, Texas (Latitude 32°37'43", Longitude 101°17'06" NAD27) provided monthly streamflow records from March 1988 until May 2021.

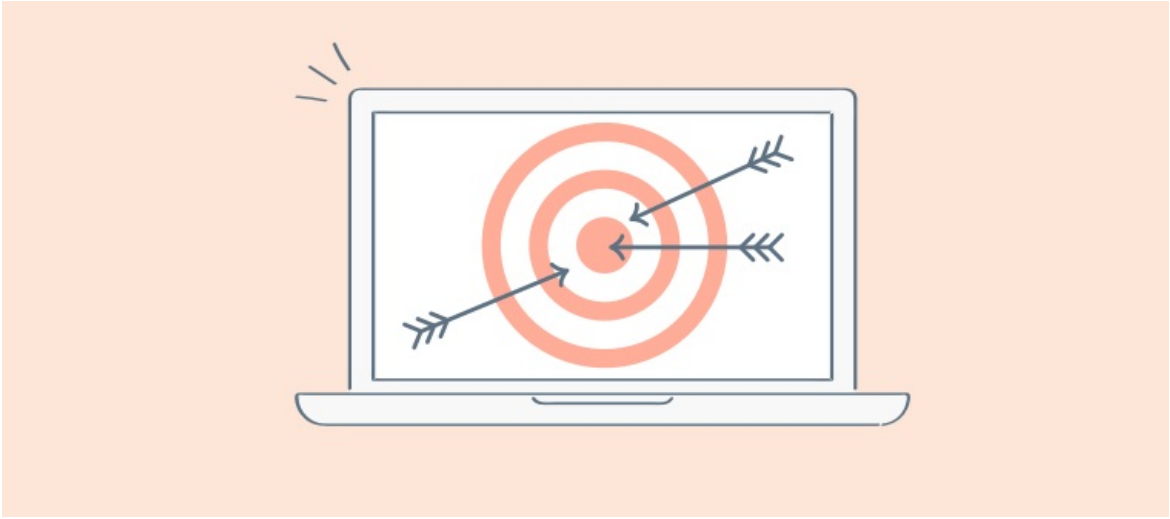
The Dataset | Meteorological Data

Meteorological data is downloaded for the same period of time and the same time scale from "prism.oregonstate.edu". This data includes monthly precipitation and temperature (max, mean, and min) records.

The Dataset | Overview

The dataset, "ColoradoRiverData.csv" is provided to you with the following information:

Columns	Info.
Date	The date of a measurement in YYYY-MM format
ppt (inches)	The total recorded precipitation in inches for each month
tmin (degrees F)	The minimum recorded temperature in degrees Fahrenheit for each month
tmean (degrees F)	The average recorded temperature in degrees Fahrenheit for each month
tmax (degrees F)	The maximum recorded temperature in degrees Fahrenheit for each month
Flowrate (cfs)	The average recorded streamflow in cubic feet per second for each month



Objective(s):

- Literature scan on the importance and approaches to streamflow forecasting.
- Analyse an existing hydro-meteorological database and build a data model to predict the streamflow based on various features.
- Build an interface to allow users to enter relative input features (e.g., precipitation) and return an estimated flowrate and an assessment of the uncertainty in the estimate
- Build an interface to allow users to enter relative input features (e.g., precipitation) and return an estimated flow state.

Tasks:

Literature Research:

In a short essay (1-2 pages):

- Describe the importance and challenges of streamflow forecasting.
- Summarize the value of a data model in the context of the conventional approach to streamflow forecasting

Some places to start are:

- Cuo, L., Pagano, T. C., & Wang, Q. J. (2011). A Review of Quantitative Precipitation Forecasts and Their Use in Short- to Medium-Range Streamflow Forecasting, Journal of Hydrometeorology, 12(5), 713-728. Retrieved Oct 21, 2021, from https://journals.ametsoc.org/view/journals/hydr/12/5/2011jhm1347_ (https://journals.ametsoc.org/view/journals/hydr/12/5/2011jhm1347_)
- Yaseen, Z. M., El-Shafie, A., Jaafar, O., Afan, H. A., & Sayl, K. N. (2015). Artificial intelligence based models for stream-flow forecasting: 2000–2015. Journal of Hydrology, 530, 829-844. available at https://www.sciencedirect.com/science/article/abs/pii/S0022169415008069 (https://www.sciencedirect.com/science/article/abs/pii/S0022169415008069)
- Zhenghao Zhang, Qiang Zhang & Vijay P. Singh (2018) Univariate streamflow forecasting using commonly used data-driven models: literature review and case study, Hydrological Sciences Journal, 63:7, 1091-1111, DOI: 10.1080/02626667.2018.1469756 available at https://www.tandfonline.com/doi/full/10.1080/02626667.2018.1469756 (https://www.tandfonline.com/doi/full/10.1080/02626667.2018.1469756)
- Mosavi A, Ozturk P, Chau K-w. Flood Prediction Using Machine Learning Models: Literature Review. Water. 2018; 10(11):1536. https://doi.org/10.3390/w10111536 (https://doi.org/10.3390/w10111536) available at https://www.mdpi.com/2073-4441/10/11/1536 (https://www.mdpi.com/2073-4441/10/11/1536)
- Cheng, M., Fang, F., Kinouchi, T., Navon, I. M., & Pain, C. C. (2020). Long lead-time daily and monthly streamflow forecasting using machine learning methods. Journal of Hydrology, 590, 125376. available at https://www.sciencedirect.com/science/article/abs/pii/S0022169420308362 (https://www.sciencedirect.com/science/article/abs/pii/S0022169420308362)

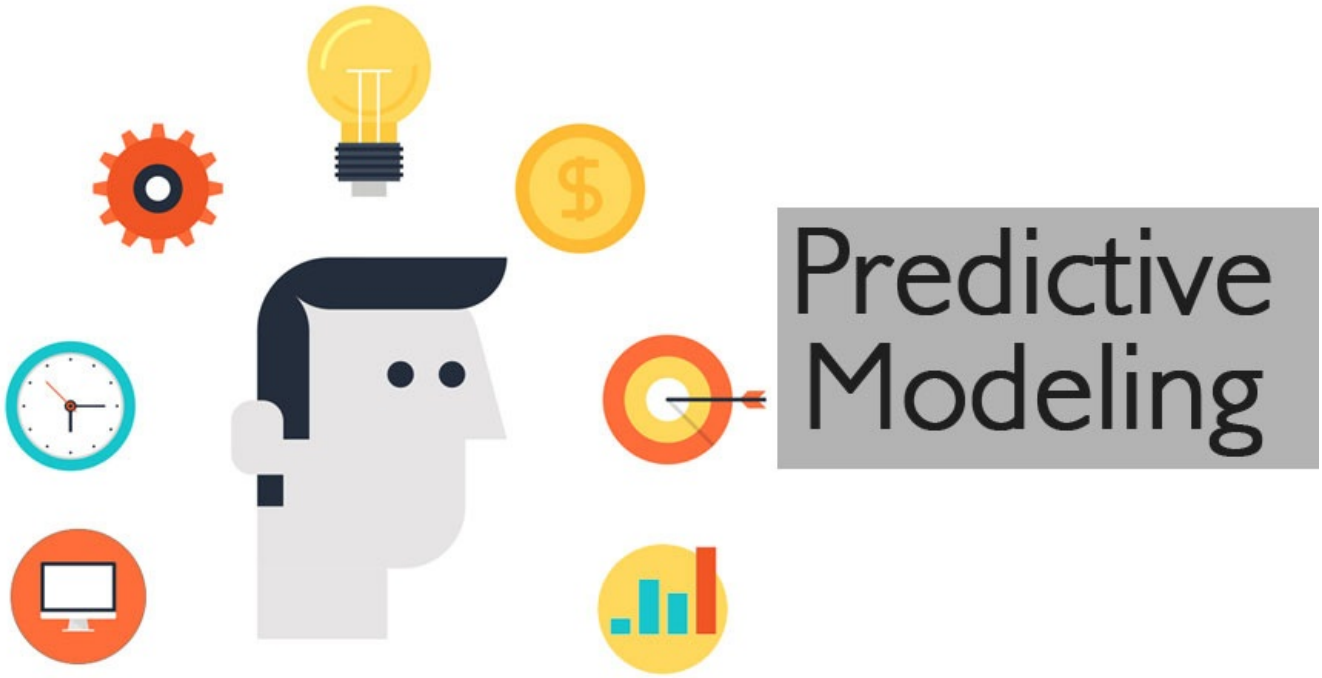
Exploratory Data Analysis (EDA):

Provide a summery (description) of the dataset in 2-3 pages. This summery should appropriately present the essential infromation about the dataset in a concise, well-written and clear manner. Things you may want to include ...

- An overall description of the dataset
- A summery of the information extracted from the important statistics of different parameters in the dataset
- A summery of the distributional charactristics of different parameters in the dataset
- A summary of the relationship status within the different parameters in the dataset

Your EDA section must include your answers for the following questions:

- Which parameter (between precipitation and temperature values) can be a better predictor for streamflow at the station of study? why?
- What can you infer from comparing the maximum recorded floods in the 90s, 2000s, and 2010s? Based on the records available in the dataset, would you expect to see more or less extreme floods in the future?
- Periods with the flowrate of 0 (No-flow periods) can be viewed as indicators of drought. What can you understand from comparing the number of no-flow days in the 90s, 2000s, and 2010s? Is the upstream of the Colorado River becoming more or less prone to drought?



Model Building: Part 1 | Forecasting flowrates

In this part, the goal is to make models to predict flowrates in the Colorado River, and then evaluate their performance using appropriate goodness-of-fit measures, and analyze the outcomes. Use the first 75% of the dataset for training your models and the remaining 25% for testing.

- Build 3 data models that you see appropriate for this task. (*Please note that these models should be unique and this uniqueness can be defined based on using different algorithms, inputs, or both.*)
- Assess data model quality (decide which model is best)
- Build the input data interface for using the "best" model
- Using your best model determine the estimated flowrate for the hydro-meteorological conditions in the table below:

ppt	tmax	tmean	tmin	last month flowrate
0.0	113.0	99.0	85.0	0.0
4.5	95.0	85.0	75.0	74.5
2.2	20.0	10.0	0.0	55.0
1.0	80.0	60.0	40.0	36.3
0.0	80.0	60.0	40.0	12.0

note that you may not all the values for each case, depending on your best model.

Your "Model Building: Part 1" section must include your answers for the following questions:

- What are the most important assumptions in your modeling?
- Is it benefitial to use the streamflow recorded in the previous step (a lagged streamflow value) as an input feature? why?
- Which parameter (between precipitation and temperature values) was be a better predictor for streamflow at the station of study? why?
- Is there a specific range of streamflow values that are harder to capture accurately for your data models? If yes, what range and why?

Model Building: Part 2 | Forecasting flow states

In this part, the goal is to make models to predict whether the Colorado River's flowstate is in the "Flow" or the "No-Flow" state. Then, evaluate their performance using appropriate goodness-of-fit measures, and analyze the outcomes. Use the first 75% of the dataset for training your models and the remaining 25% for testing.

- Add a column to the dataframe for the flow state: It should be 0 when the flowrate is equal to 0 and 1 when the flowrate is non-zero.
- Build 3 data models that you see appropriate for predicting the flow state. (*Please note that these models should be unique and this uniqueness can be defined based on using different algorithms, inputs, or both.*)
- Assess data model quality (decide which model is best)
- Build the input data interface for using the "best" model
- Using your best model determine the estimated flow state for the hydro-meteorological conditions in the table below:

ppt	tmax	tmean	tmin
0.0	113.0	99.0	85.0
4.5	95.0	85.0	75.0

2.2	20.0	10.0	0.0
1.0	80.0	60.0	40.0
0.0	80.0	60.0	40.0

note that you may not all the values for each case, depending on your best model.

Your "Model Building: Part 2" section must include your answers for the following questions:

- What are the most important assumptions in your modeling?
- Is it more difficult for the model to capture one flow state more than the other? If yes, which one? why?
- Which parameter (between precipitation and temperature values) was be a better predictor for flow state at the station of study? why?
- What can be some applications of this kind of streamflow state modeling?

Deliverables:

Effort Sheets (due every week on Friday):

Each team must submit an effort sheet which is a table with a clear discription of the tasks undertaken by each member and has the signiture of all team members. The effort sheets should be submitted digitally via email.

Interim report (due November 24):

This report must include:

- The "Literature Research" section
- A description of the Colorado River database
- A plan of work for how you want to handle the project and solve the modeling tasks.
- Break down each task into manageable subtasks and describe how you intend to solve the subtasks and how you will test each task. (Perhaps make a simple Gantt Chart)
- Address the responsibilities of each team member for tasks completed and tasks to be completed until the end of the semester. (Perhaps make explicit subtask assignments)

Your report should be limited to 7 pages, 12 pt font size, double linespacing (exclusive of references which are NOT included in the page count). You need to cite/reference all sources you used. This report must be submitted by Midnight November 24th in PDF format.

Final report (due December 7):

This report must include:

- All the required parts, including the ones from the Interim report as well as the sections on EDA and Model Building parts.
- All the filled effort sheets with the signitures of the team members with a clear description of all the tasks performed by each member.
- All the references used in the entire length of the project.
  This report must be submitted by Midnight December 7th in PDF format, along with the following documents:
  - A well-documented Jupyter Notebook (.ipynb file) for the analysis and implementation of the data models.
  - A well-documented Jupyter Notebook (.ipynb file) for the implementation of the data model user interface.

    Above items can reside in a single notebook; but clearly identify sections that perform different tasks.
  - A how-to video demonstrating the application, performance and description of what you did for the project, including the problems that you solved as well as those that you were not able to solve.
  - A project management video (up to 5 minutes) in which you explain how you completed the project and how you worked as a team.

    Above items can reside in a single video; but structure the video into the two parts; use an obvious transition when moving from "how to ..." into the project management portion.
    Keep the total video length to less than 10 minutes; submit as an "unlisted" YouTube video, and just supply the link (someone on each team is likely to have a YouTube creator account). Keep in mind a 10 minute video can approach 100MB file size before compression, so it won't upload to Blackboard and cannot be emailed.


HAPPY
CODING