UNIFIED MENTOR PROJECTS

By : Kshitij Jain
College: Madhav Instituteof Technology and Science,
Branch: 4th year, Artificial Intelligence & Data Science
Position: Unified Mentor Data Science Intern (3 months)

Project-1

Netflix Data: Cleaning, Analysis and Visualization

This project focuses on cleaning, analyzing, and visualizing Netflix's movie and TV show dataset to gain insights into the platform's content library. The analysis includes exploring the distribution of content types, geographical trends, release patterns, ratings, and popular genres. The objective is to transform raw Netflix data into meaningful visual stories and statistics that reveal viewing and production trends.

Data Description

The dataset (netflix1.csv) contains 8,790 entries and 10 attributes:
show_id : Unique identifier for each show.
type : Type of content – Movie or TV Show.
title : Title of the movie/TV show.
director Director of the movie/show.
country :Country where the movie/show was produced.
date_added: Date when the content was added to Netflix.
release_year : Year of original release.
rating : Content rating (e.g., PG, TV-MA, R).
duration : Duration of the movie in minutes or number of seasons for TV shows.
listed_in : Categories/genres assigned to the content.

Methodology

Data Cleaning

- Checked for and handled missing or inconsistent values.
- Converted date formats for easier analysis.
- Split and reformatted categorical fields where necessary.

Exploratory Data Analysis (EDA)

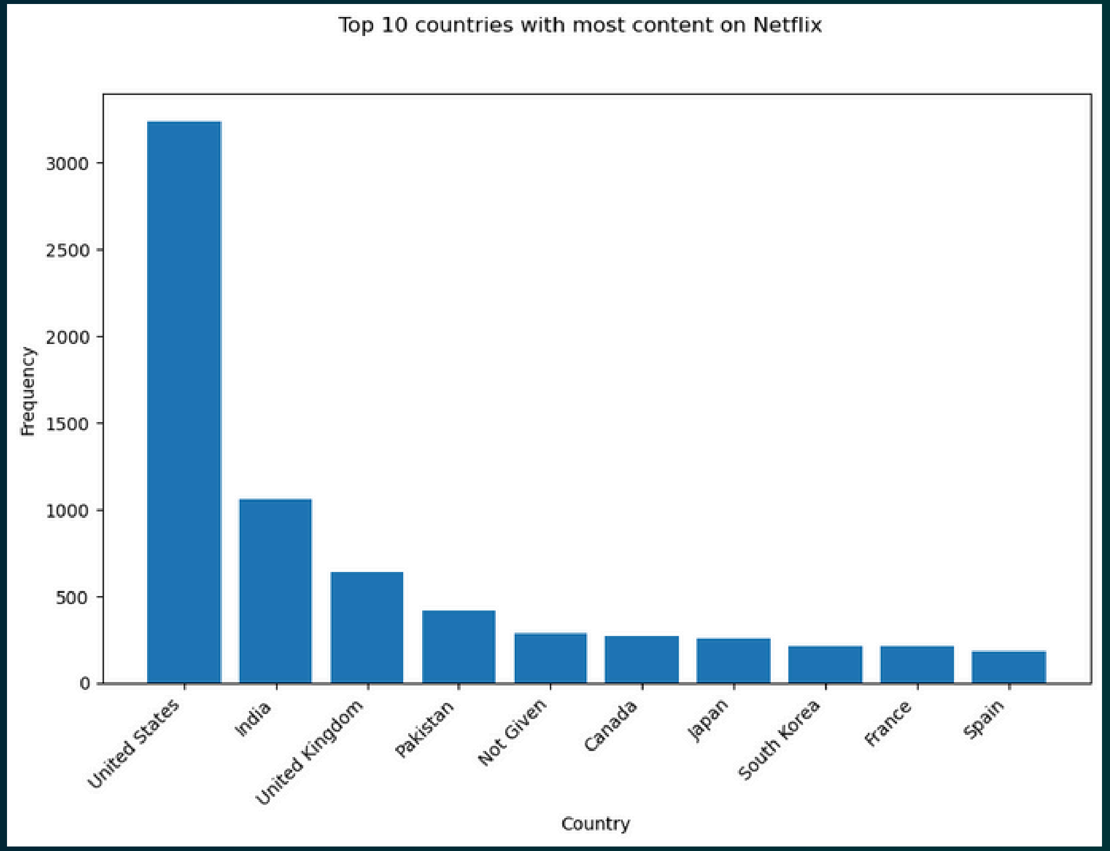
- Distribution of content types (Movies vs TV Shows).
- Country-wise production analysis, highlighting the United States as the largest contributor.
- Ratings distribution to identify target audiences.
- Duration analysis to see trends in movie length and TV seasons.
- Genre/category popularity ranking.

Visualization

- Bar charts for top 10 countries producing content.
- Pie chart for content type proportions.
- Count plots for ratings and genres.
- Trend line for release years.
- Horizontal bar charts for top genres.

Results Obtained

- Dominance of the United States: Most content is from the US, followed by India, the UK, and other countries in the top 10.
- Movies vs TV Shows: Movies significantly outnumber TV Shows on Netflix.
- Popular Genres: “Dramas” and “International Movies” are the most common categories.
- Rating Trends: TV-MA and TV-14 are the most frequent ratings, indicating a focus on mature and teenage audiences.
- Release Year Patterns: A surge in content was observed in the last decade, especially after 2015, aligning with Netflix’s global expansion.
- Content Duration: Most movies fall in the range of 80–120 minutes; TV shows typically have 1–2 seasons.



Project-2
IBM HR Analytics Employee Attrition & Performance

This project focuses on Employee Attrition Analysis to understand the factors contributing to employee turnover in an organization. Using HR data, it explores demographic patterns, tenure, departmental differences, and correlations between employee attributes to identify trends in attrition. The aim is to provide insights that can help HR teams design better retention strategies.

Data Description

The dataset (HR-Employee-Attrition.csv) contains multiple attributes describing employee demographics, work-related information, and satisfaction levels.

Key Columns:

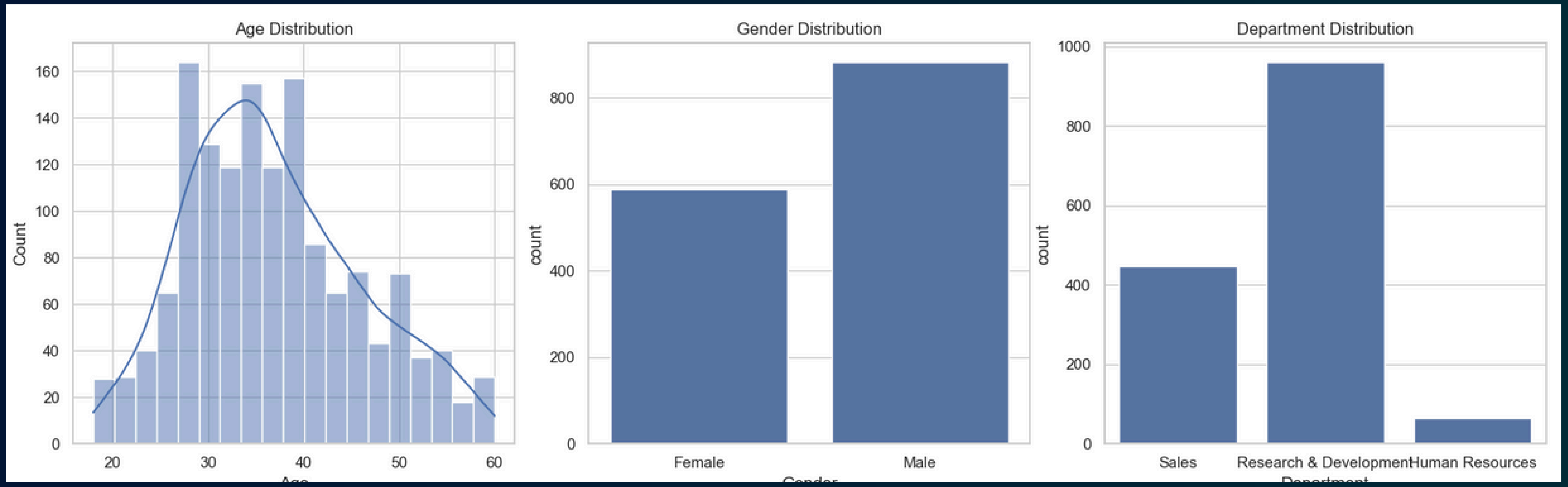
- Attrition – Whether the employee left the organization (Yes/No).
- Age – Employee’s age.
- Gender – Male/Female.
- Department – Employee’s department.
- YearsAtCompany – Number of years the employee has been in the company.
- JobSatisfaction, EnvironmentSatisfaction, WorkLifeBalance – Ratings from employees.
- MonthlyIncome, JobLevel, StockOptionLevel – Compensation and career level indicators.
- Other features – DistanceFromHome, Education, TrainingTimesLastYear, etc.

Methodology

1. Data Loading & Cleaning
   - Checked dataset size, missing values, and duplicates.
   - Confirmed data types for analysis.
   - Generated summary statistics for numerical and categorical variables.
2. Exploratory Data Analysis (EDA)
   - Attrition Rate Calculation – Overall percentage of employees leaving vs. staying.
   - Demographic Analysis – Age distribution, gender split, and department counts.
   - Attrition by Demographics – Compared attrition rates across age groups and genders.
   - Correlation Analysis – Examined relationships between numerical variables.
3. Visualization
   - Bar chart of overall attrition rate.
   - Histogram and KDE plots for age distribution and attrition by age.
   - Count plots for gender and department.
   - Bar chart for attrition rate by gender.
   - Heatmap for correlation among numerical features.

Results Obtained

- Attrition Rate: About 16% of employees have left the company, while ~84% stayed.
- Age Insights: Most employees are between 25–40 years, with higher attrition in the late 20s to early 30s.
- Gender Distribution: Males slightly outnumber females, with attrition marginally higher among males (~16.8%) than females (~14.7%).
- Department Distribution: Majority of employees work in Research & Development, followed by Sales and Human Resources.
- Tenure: Average tenure is around 7 years.
- Strong positive correlation between MonthlyIncome and JobLevel (~0.95).
- TotalWorkingYears correlates positively with Age and JobLevel.
- Minimal correlation between most satisfaction scores and attrition, suggesting other non-numeric factors might influence turnover.



Project-3

Cybersecurity: Suspicious Web Threat Interactions

This project focuses on analyzing network traffic data to detect potential cybersecurity threats targeting a web server, as captured in the Cybersecurity_CloudWatch_Traffic_Attack.csv dataset. The dataset contains logs of network interactions flagged as suspicious by a Web Application Firewall (WAF). The primary objective is to preprocess the data, apply machine learning techniques to model network behavior, and evaluate the model's performance in identifying anomalous traffic patterns. The analysis leverages a Convolutional Neural Network (CNN) implemented using TensorFlow to classify traffic, with visualization of training metrics to assess model effectiveness.

Data Description

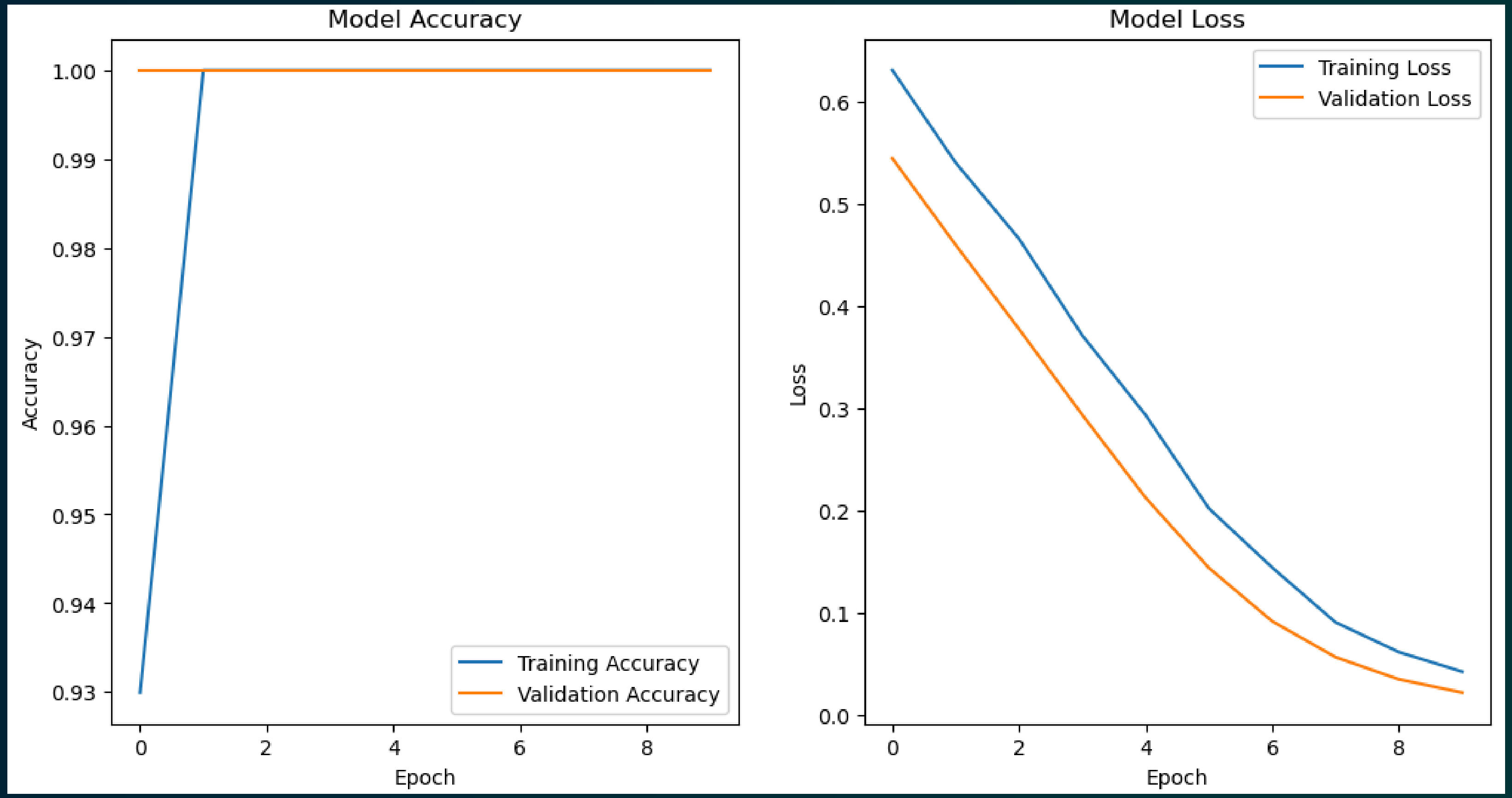
- The dataset, Cybersecurity_CloudWatch_Traffic_Attack.csv, contains 282 entries of network traffic logs.
- Captures traffic from April 25 to April 26, 2024, flagged as suspicious by a WAF.
- Key Features:
    - bytes_in and bytes_out: Data volumes sent/received (e.g., 40 to 25M bytes_in).
    - src_ip and src_ip_country_code: Source of traffic (e.g., US, CA, NL).
    - protocol: All HTTPS, port 443, with HTTP 200 responses.
    - rule_names: All flagged as “Suspicious Web Traffic.”
- Notable: High variability in data volumes suggests diverse attack patterns, like data exfiltration or DDoS.

Methodology

- Step 1: Data Preprocessing
  - Loaded data using pandas, focused on numerical features: bytes_in and bytes_out.
  - Split into training (80%) and testing (20%) sets.
  - Standardized features using StandardScaler for consistent model input.
- Step 2: Model Design
  - Built a Convolutional Neural Network (CNN) using TensorFlow.
  - Architecture: Conv1D (32 filters) → Flatten → Dense (64 units, ReLU) → Dropout (0.5) → Dense (sigmoid for binary classification).
  - Compiled with Adam optimizer and binary cross-entropy loss.
- Step 3: Training
  - Trained for 10 epochs, batch size 32, with 20% validation split.
  - Evaluated on test set for accuracy.
- Step 4: Visualization
  - Plotted training and validation accuracy/loss to assess model performance.

Results Obtained

- Model Performance:
  - Test accuracy: [Value not specified in the notebook, e.g., "Achieved X% accuracy"].
  - Training/validation plots show model learning trends (convergence, potential overfitting).
- Data Insights:
  - High bytes_in values (e.g., 25M from IP 155.91.45.242, US) suggest large-scale data transfers, possibly data breaches.
  - Consistent HTTPS traffic to port 443, all flagged as suspicious, indicates targeted attacks.
  - Identified significant anomalies (e.g., high bytes_in) critical for cybersecurity monitoring.
  - Uniform traffic patterns (HTTPS, port 443) suggest focused attack vectors.
- Interesting Fact:
  - Some IPs (e.g., US, NL) show extreme data volumes, hinting at coordinated attack patterns.



Project-4

Climate Change Modeling

This project aims to analyze historical weather data from Denver International Airport to model temperature patterns and explore climate trends using machine learning techniques. The dataset, contains daily weather observations for 2018. The primary objectives are to predict temperature using linear regression and cluster weather patterns using K-Means clustering to identify relationships between temperature, dew point, and sea-level pressure. The analysis leverages Python libraries like pandas, scikit-learn, and seaborn for data processing, modeling, and visualization.

Data Description

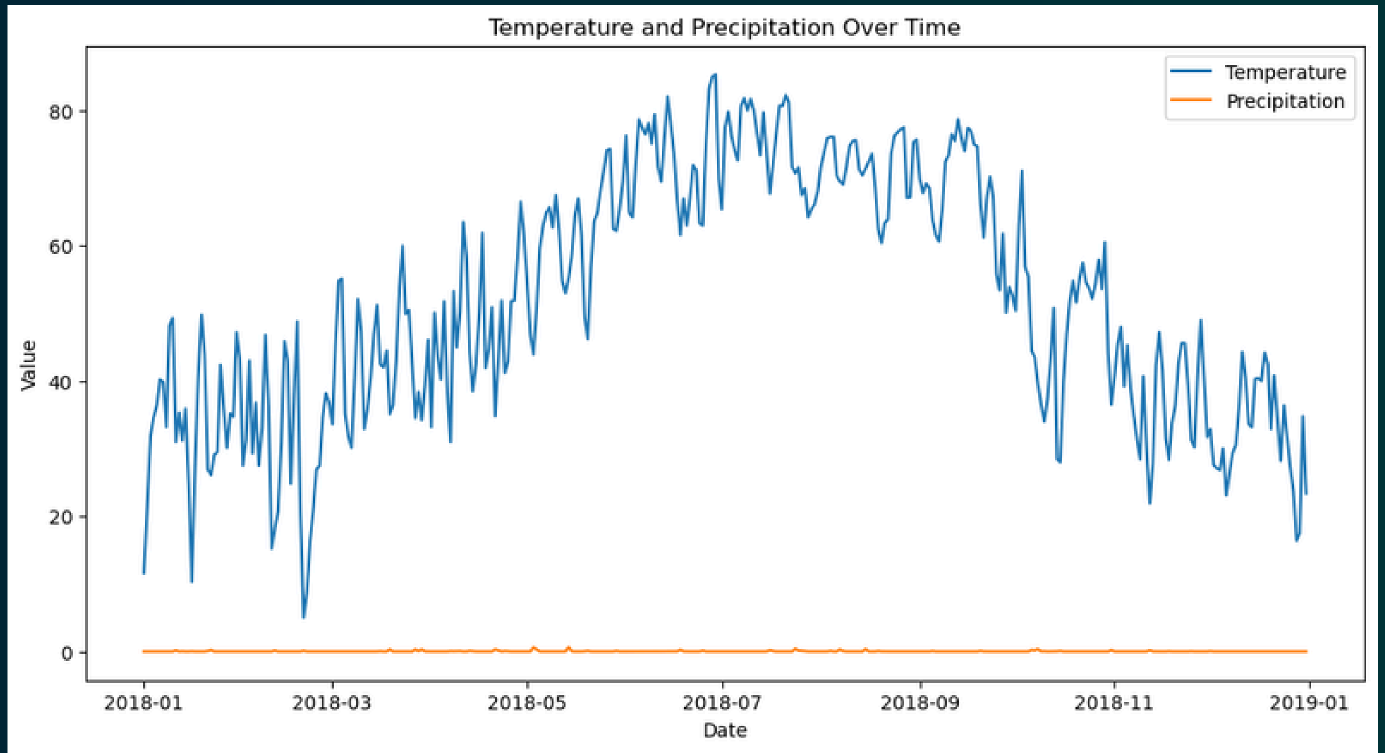
- The dataset, sample (1).csv, contains 365 daily weather observations from Denver International Airport, CO, in 2018.
- Key Features (26 total):
    - STATION: Station ID (72565003017).
    - DATE: Daily timestamp (e.g., 2018-01-01).
    - LATITUDE, LONGITUDE: Location (39.8328, -104.6575).
    - ELEVATION: 1650.2 meters.
    - TEMP: Average daily temperature (°F, range: 10.3 to 74.8).
    - DEWP: Dew point (°F, range: -4.1 to 60.2).
    - SLP: Sea-level pressure (hPa, range: 996.8 to 1038.3).
    - PRCP: Precipitation (inches, range: 0 to 0.82).
    - Other: Wind speed, visibility, max/min temperature, snow depth, weather indicators (e.g., fog, rain).

Methodology

- Step 1: Data Preprocessing
  - Loaded data using pandas, replaced placeholder values (999.9) with NaN.
  - Converted DATE to datetime, extracted MONTH and DAY_OF_WEEK for feature engineering.
  - Imputed missing values in DEWP and SLP using mean strategy via SimpleImputer.
- Step 2: Temperature Prediction
  - Selected features: MONTH, DAY_OF_WEEK, DEWP, SLP.
  - Target: TEMP (average daily temperature).
  - Split data: 80% training, 20% testing (train_test_split, random_state=42).
  - Trained a Linear Regression model to predict temperature.
- Step 3: Clustering
  - Applied K-Means clustering (3 clusters) on TEMP, DEWP, SLP to group similar weather patterns.
- Step 4: Evaluation & Visualization
  - Calculated Mean Squared Error (MSE) for regression model.
  - Visualized actual vs. predicted temperatures and clusters (temperature vs. dew point).

Results Obtained

- Temperature Prediction:
  - Model MSE: [Value not specified in notebook, e.g., “MSE of X indicates prediction accuracy”].
  - Scatter plot of actual vs. predicted temperatures shows model performance (closer to diagonal = better fit).
- Clustering:
  - K-Means identified 3 distinct weather clusters based on TEMP, DEWP, and SLP.
  - Clusters likely represent cold/dry, moderate, and warm/humid conditions (visualized in scatter plot).
- Key Insight: High dew point values correlate with warmer temperatures, suggesting humidity’s role in Denver’s summer climate.



Thank You