DSL251
Data Analytics and Visualization
Homework 1

Amay Dixit
12340220

January 19, 2025

Question 1

Problem Statement:

Given a uniform distribution of integers $n \in \{1, 2, \ldots, 1000\}$, let x be the first digit of n. We are tasked with:

1. Compute the probabilities $p_1, p_2, \ldots, p_9$ of $x = 1, 2, \ldots, 9$, respectively.
2. Determine the mean $\mathbb{E}[x]$ and variance $\mathrm{Var}(x)$ of x.

Solution

Step 1: Determining the Range of x

The value of x depends on the range of n. For $x \in \{1, 2, \ldots, 9\}$, we compute the number of integers that start with each digit x within the range $n \in \{1, 2, \ldots, 1000\}$.

- For $x = 1$:
  - Numbers in the range $[1, 9]$: 1 number.
  - Numbers in the range $[10, 19]$: 10 numbers.
  - Numbers in the range $[100, 199]$: 100 numbers.
  - Number 1000: 1 number.

  Total: $1 + 10 + 100 + 1 = 112$.

- For $x = 2, 3, \ldots, 9$:
  - Numbers in the range $[1, 9]$: 1 number.
  - Numbers in the range $[x \cdot 10, x \cdot 10 + 9]$: 10 numbers.
  - Numbers in the range $[x \cdot 100, x \cdot 100 + 99]$: 100 numbers.

  Total: $1 + 10 + 100 = 111$ for each $x \in \{2, 3, \ldots, 9\}$.

Step 2: Computing the Probabilities p_x

The probability p_x of x being the first digit is:

$$p_x = \frac{\text{Count of numbers starting with } x}{1000}.$$

For $x = 1$:

$$p_1 = \frac{112}{1000} = 0.112.$$

For $x = 2, 3, \ldots, 9$:

$$p_x = \frac{111}{1000} = 0.111.$$

Thus, the probabilities are:

$$p_1 = 0.112, \quad p_2 = p_3 = \cdots = p_9 = 0.111.$$

Step 3: Computing the Mean $\mathbb{E}[x]$

The mean $\mathbb{E}[x]$ is given by:

$$\mathbb{E}[x] = \sum_{x=1}^{9} x \cdot p_x.$$

Substituting the values of p_x:

$$\mathbb{E}[x] = 1 \cdot 0.112 + 2 \cdot 0.111 + 3 \cdot 0.111 + \cdots + 9 \cdot 0.111.$$

Simplifying the summation:

$$\mathbb{E}[x] = 0.112 + 0.222 + 0.333 + 0.444 + 0.555 + 0.666 + 0.777 + 0.888 + 0.999 = 4.996.$$

Step 4: Computing the Variance $\mathrm{Var}(x)$

The variance $\mathrm{Var}(x)$ is given by:

$$\mathrm{Var}(x) = \mathbb{E}[x^2] - (\mathbb{E}[x])^2,$$

where:

$$\mathbb{E}[x^2] = \sum_{x=1}^{9} x^2 \cdot p_x.$$

To compute $\mathbb{E}[x^2]$, we have:

$$\mathbb{E}[x^2] = 1^2 \cdot 0.112 + 2^2 \cdot 0.111 + 3^2 \cdot 0.111 + \cdots + 9^2 \cdot 0.111.$$

Substituting the values:

$$\mathbb{E}[x^2] = 1 \cdot 0.112 + 4 \cdot 0.111 + 9 \cdot 0.111 + 16 \cdot 0.111 + 25 \cdot 0.111 + 36 \cdot 0.111 + 49 \cdot 0.111 + 64 \cdot 0.111 + 81 \cdot 0.111.$$

$$\mathbb{E}[x^2] = 31.636.$$

Now, compute the variance:

$$\mathrm{Var}(x) = 31.636 - (4.996)^2 = 6.675984.$$

Question 2

Given n identical Gaussian distributions, each with $N(0, \sigma^2)$, we are tasked with finding the variance of their convolution.

Gaussian Distribution Convolution

A Gaussian distribution with mean 0 and variance σ^2 is given by the probability density function (PDF):

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}.$$

Represented as $X \sim N(0, \sigma^2)$ with mean 0 and variance σ^2

For Gaussian functions, the convolution of two Gaussian distributions with means $\mu_1 = \mu_2 = 0$ and variances σ_1^2 and σ_2^2 is still a Gaussian function, with the mean being the sum of the means, and the variance being the sum of the variances. That is, if:

$$p_1(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{x^2}{2\sigma_1^2}}, \quad p_2(x) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{x^2}{2\sigma_2^2}},$$

then the convolution $(p_1 * p_2)(x)$ is given by:

$$(p_1 * p_2)(x) = \frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}} e^{-\frac{x^2}{2(\sigma_1^2+\sigma_2^2)}}.$$

and if $\sigma_1^2 = \sigma_2^2 = \sigma^2$, the convolution simplifies to:

$$(p_1 * p_2)(x) = \frac{1}{\sqrt{2\pi(2\sigma^2)}} e^{-\frac{x^2}{2(2\sigma^2)}}.$$

Thus, the result of convolving two identical Gaussians is another Gaussian with mean 0 and variance $2\sigma^2$.

Generalization to n Identical Gaussians

The generalization of this result to the convolution of n identical Gaussian distributions can be given using the property that the convolution of n identical Gaussian distributions $p(x)$, each with variance σ^2, is a Gaussian distribution with mean 0 and variance equal to the sum of the individual variances. Therefore, the convolution of n identical Gaussians is given by:

$$(p_1 * p_2 * \cdots * p_n)(x) = \frac{1}{\sqrt{2\pi(n\sigma^2)}} e^{-\frac{x^2}{2(n\sigma^2)}}.$$

Thus, the resulting distribution after convolving n identical Gaussians has variance $n\sigma^2$.

Alternative Fourier Transform Method

Alternatively, we can compute the convolution using the Fourier transform method as mentioned in the question. The Fourier transform of the convolution of two functions is the product of their individual Fourier transforms, i.e.,

$$\mathcal{F}[p_1 * p_2] = \mathcal{F}(p_1) \cdot \mathcal{F}(p_2).$$

For a Gaussian distribution $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$, the Fourier transform is given by:

$$\mathcal{F}(p(x))(k) = e^{-\frac{\sigma^2 k^2}{2}}.$$

Therefore, the Fourier transform of the convolution of n identical Gaussians is:

$$\mathcal{F}[p_1 * p_2 * \cdots * p_n](k) = \left(e^{-\frac{\sigma^2 k^2}{2}}\right)^n = e^{-\frac{n\sigma^2 k^2}{2}}.$$

Taking the inverse Fourier transform of this result gives:

$$\mathcal{F}^{-1}\left(e^{-\frac{n\sigma^2 k^2}{2}}\right) = \frac{1}{\sqrt{2\pi(n\sigma^2)}} e^{-\frac{x^2}{2(n\sigma^2)}}.$$

Thus, the convolution of n identical Gaussian distributions results in a Gaussian distribution with variance $n\sigma^2$.

Question 3

Let X and Y be two independent random variables with respective probability mass functions (PMFs) $p_X(x)$ and $p_Y(y)$. We need to find the probability mass function $P_Z(z)$ of the random variable $Z = X + Y$.

Now, we know

$$P_Z(z) = \mathbb{P}(Z = z) = \mathbb{P}(X + Y = z).$$

To compute this, we consider all possible pairs of values (x, y) such that $x + y = z$. Since X and Y are independent, the joint probability, the following can be written:

$$\mathbb{P}(X = x \text{ and } Y = y) = p_X(x) \cdot p_Y(y).$$

Thus, the total probability $P_Z(z)$ is obtained by summing over all such pairs (x, y):

$$P_Z(z) = \sum_x \mathbb{P}(X = x \text{ and } Y = z - x) = \sum_x p_X(x) \cdot p_Y(z - x).$$

Also, the convolution operation $(p_X * p_Y)(z)$ is defined as:

$$(p_X * p_Y)(z) = \sum_x p_X(x) \cdot p_Y(z - x).$$

Which is equal to what we got above, thus the probability distribution of $Z = X + Y$ is given by:

$$P_Z(z) = (p_X * p_Y)(z).$$

Example: Rolling Two Dice

Let us consider two random variable X and Y representing the outcomes of rolling two six-sided dice. The PMF for each die is:

$$p_X(x) = \begin{cases} \frac{1}{6}, & \text{if } x \in \{1, 2, 3, 4, 5, 6\}, \\ 0, & \text{otherwise.} \end{cases}$$

The PMF $P_Z(z)$ for $Z = X + Y$, the sum of the dice, is obtained by computing the convolution $(p_X * p_Y)(z)$. Let us calculate $P_Z(z)$ for specific values of z:

- For $z = 2$:
$$P_Z(2) = \sum_x p_X(x) \cdot p_Y(2 - x).$$
Here, the only valid pair is $(x = 1, y = 1)$. Thus:
$$P_Z(2) = p_X(1) \cdot p_Y(1) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}.$$

- For $z = 3$:
$$P_Z(3) = \sum_x p_X(x) \cdot p_Y(3 - x).$$
The valid pairs are $(x = 1, y = 2)$ and $(x = 2, y = 1)$. Thus:
$$P_Z(3) = p_X(1) \cdot p_Y(2) + p_X(2) \cdot p_Y(1) = \frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} = \frac{2}{36}.$$

- For $z = 4$:
$$P_Z(4) = \sum_x p_X(x) \cdot p_Y(4 - x).$$
The valid pairs are $(x = 1, y = 3), (x = 2, y = 2), (x = 3, y = 1)$. Thus:
$$P_Z(4) = p_X(1) \cdot p_Y(3) + p_X(2) \cdot p_Y(2) + p_X(3) \cdot p_Y(1) = \frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} + \frac{1}{6} \cdot \frac{1}{6} = \frac{3}{36}.$$

Thus, for $z \in \{2, 3, \ldots, 12\}$, we obtain the full distribution:

$$P_Z(z) = \begin{cases} \frac{1}{36}, & \text{if } z = 2 \text{ or } 12, \\ \frac{2}{36}, & \text{if } z = 3 \text{ or } 11, \\ \frac{3}{36}, & \text{if } z = 4 \text{ or } 10, \\ \frac{4}{36}, & \text{if } z = 5 \text{ or } 9, \\ \frac{5}{36}, & \text{if } z = 6 \text{ or } 8, \\ \frac{6}{36}, & \text{if } z = 7. \end{cases}$$

Thus, The probability distribution $P_Z(z)$ for the sum of two independent random variables X and Y is given by the convolution of their individual distributions.

Question 4

To prove the following identity, that for random samples $0 < x_1 < x_2 < \cdots < x_n$ with probabilities $p_1, p_2, \ldots, p_n$:

$$\textbf{mean} = E[x] = \sum_{i=1}^{n} p_i x_i = \int_{t=0}^{\infty} (\textbf{Probability that } X > t)\, dt.$$

We start by evaluating the RHS Integral:

$$\int_{t=0}^{\infty} P(X > t)\, dt.$$

To calculate this, split the integral as follows

$$\int_{t=0}^{\infty} P(X > t)\, dt = \int_{0}^{x_1} P(X > t)\, dt + \int_{x_1}^{x_2} P(X > t)\, dt + \cdots + \int_{x_{n-1}}^{x_n} P(X > t)\, dt + \int_{x_n}^{\infty} P(X > t)\, dt.$$

Evaluating $P(X > t)$

For each interval, the probability $P(X > t)$ takes on constant values:

- For $t \in [0, x_1]$: All values exceed t, so $P(X > t) = 1$.
- For $t \in [x_1, x_2]$: $P(X > t) = \sum_{i=2}^{n} p_i$ (probability of values greater than x_1).
- For $t \in [x_2, x_3]$: $P(X > t) = \sum_{i=3}^{n} p_i$.
- And so on, until:
- For $t \in [x_n, \infty]$: $P(X > t) = 0$ (no values exceed x_n).

Computing the Integral

Substituting these values into the integral, we get:

$$\int_{t=0}^{\infty} P(X > t)\, dt = \int_{0}^{x_1} 1\, dt + \int_{x_1}^{x_2} \sum_{i=2}^{n} p_i\, dt + \cdots + \int_{x_{n-1}}^{x_n} p_n\, dt.$$

Evaluating each term:

$$\int_{0}^{x_1} 1\, dt = x_1,$$

$$\int_{x_1}^{x_2} \sum_{i=2}^{n} p_i\, dt = (x_2 - x_1) \sum_{i=2}^{n} p_i,$$

$$\int_{x_2}^{x_3} \sum_{i=3}^{n} p_i\, dt = (x_3 - x_2) \sum_{i=3}^{n} p_i,$$

and so on, until:

$$\int_{x_{n-1}}^{x_n} p_n\, dt = (x_n - x_{n-1}) p_n.$$

Thus, the total integral becomes:

$$\int_{t=0}^{\infty} P(X>t)\,dt = x_1 + (x_2 - x_1)\sum_{i=2}^{n} p_i + (x_3 - x_2)\sum_{i=3}^{n} p_i + \cdots + (x_n - x_{n-1})p_n.$$

Simplifying,

$$\int_{t=0}^{\infty} P(X>t)\,dt = p_1x_1 + p_2x_2 + p_3x_3 + \cdots + p_nx_n = \sum_{i=1}^{n} p_ix_i.$$

Also,

$$E[X] = \sum_{i=1}^{n} p_ix_i.$$

Thus,

$$E[X] = \int_{t=0}^{\infty} P(X>t)\,dt.$$

Hence Proved

Question 5

For the given Gaussian distribution,

$$0.4\mathcal{N}\left(\begin{bmatrix}10\\2\end{bmatrix},\ \begin{bmatrix}1&0\\0&1\end{bmatrix}\right)+0.6\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\ \begin{bmatrix}8.4&2.0\\2.0&1.7\end{bmatrix}\right).$$

The first component has mean vector $\mu_1 = \begin{bmatrix}10\\2\end{bmatrix}$ and covariance matrix $\Sigma_1 = \begin{bmatrix}1&0\\0&1\end{bmatrix}$, while the second component has mean vector $\mu_2 = \begin{bmatrix}0\\0\end{bmatrix}$ and covariance matrix $\Sigma_2 = \begin{bmatrix}8.4&2.0\\2.0&1.7\end{bmatrix}$. The weights are 0.4 for the first component and 0.6 for the second.

(a) Compute the marginal distributions for each dimension

The marginal distributions for each dimension can be obtained by integrating the other dimensions in the joint distribution.

For the first dimension:

The marginal distribution of the first dimension X_1 is as follows,

$$X_1 \sim 0.4\mathcal{N}(10,1) + 0.6\mathcal{N}(0,8.4).$$

Thus, the mean and variance of the marginal distribution for X_1 are computed as:

$$\mu_{X_1} = 0.4\times 10 + 0.6\times 0 = 4.$$

$$\sigma^2_{X_1} = 0.4\times 1 + 0.6\times 8.4 = 5.04.$$

So, the marginal distribution for X_1 is:

$$X_1 \sim \mathcal{N}(4, 5.04).$$

For the second dimension:
Similarly, for the second dimension X_2,

$$X_2 \sim 0.4\mathcal{N}(2, 1) + 0.6\mathcal{N}(0, 1.7).$$

The mean and variance for X_2 are:

$$\mu_{X_2} = 0.4 \times 2 + 0.6 \times 0 = 0.8,$$

$$\sigma^2_{X_2} = 0.4 \times 1 + 0.6 \times 1.7 = 1.48.$$

Thus, the marginal distribution for X_2 is:

$$X_2 \sim \mathcal{N}(0.8, 1.48).$$

(b) Compute the mean, mode, and median for each marginal distribution

In a Gaussian Distribution, mean = median = mode, so
For the first dimension X_1:

- $\mu_{X_1} = mean = mode = median = 4$

For the second dimension X_2:

- $\mu_{X_1} = mean = mode = median = 0.8$

(c) Compute the mean and mode for the two-dimensional distribution

The two-dimensional distribution is a mixture of two 2D Gaussian distributions. The mean of the mixture distribution is the weighted sum of the means of the individual components:

$$\mu = 0.4 \times \begin{bmatrix} 10 \\ 2 \end{bmatrix} + 0.6 \times \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 4 \\ 0.8 \end{bmatrix}.$$

For the mode, since both components are Gaussian distributions, and the mixture weights are not symmetric, the mode of the mixture distribution will coincide with the mode of the component with the highest weight. Since the second component has a higher weight (0.6 ¿ 0.4), the mode of the mixture distribution will be the mean of the second component, which is:

$$\text{Mode} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Thus, the mean and mode for the two-dimensional distribution are:

$$\text{Mean} = \begin{bmatrix} 4 \\ 0.8 \end{bmatrix}, \quad \text{Mode} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Question 6

The likelihood function for the Bernoulli distribution is given by:

$$p(x \mid \mu) = \mu^2(1-\mu)^{1-x}(1-x), \quad x \in \{0,1\}$$

Conjugate Prior:

For the Bernoulli likelihood, the conjugate prior for μ is the Beta distribution. Thus, we assume the prior $p(\mu)$ follows a Beta distribution with parameters α and β:

$$p(\mu) = \text{Beta}(\mu \mid \alpha, \beta) = \frac{\mu^{\alpha-1}(1-\mu)^{\beta-1}}{B(\alpha,\beta)}$$

where $B(\alpha,\beta)$ is the Beta function.

Posterior Distribution:

To find the posterior distribution $p(\mu \mid x_1, \ldots, x_N)$, we use Bayes' theorem:

$$p(\mu \mid x_1, \ldots, x_N) \propto p(\mu) \prod_{i=1}^{N} p(x_i \mid \mu)$$

Substituting the likelihood $p(x_i \mid \mu)$ and the prior $p(\mu)$:

$$p(\mu \mid x_1, \ldots, x_N) \propto \left(\mu^{\alpha-1}(1-\mu)^{\beta-1}\right) \prod_{i=1}^{N} \mu^2(1-\mu)^{1-x_i}(1-x_i)$$

Simplifying the expression:

$$p(\mu \mid x_1, \ldots, x_N) \propto \mu^{2N+\alpha-1}(1-\mu)^{N-\sum_{i=1}^{N} x_i+\beta-1}$$

4. Normalizing the Posterior:

The posterior is in the form of a Beta distribution. Thus, we can write the posterior distribution as:

$$p(\mu \mid x_1, \ldots, x_N) = \text{Beta}(\mu \mid \alpha' = \alpha + 2N, \beta' = \beta + N - \sum_{i=1}^{N} x_i)$$

Final Answer:

$$p(\mu \mid x_1, \ldots, x_N) = \text{Beta}(\mu \mid \alpha + 2N, \beta + N - \sum_{i=1}^{N} x_i)$$

Question 7

We will apply Bayes' Theorem to solve the problem. Let:

- A: The event that the mango was picked from bag 2.
- B: The event that the fruit picked is a mango.

We need to find $P(A \mid B)$, the probability that the mango was picked from bag 2 given that the fruit is a mango.

By Bayes' theorem:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Substituting the events in terms of the problem:

$$P(\text{Bag 2} \mid \text{Mango}) = \frac{P(\text{Mango} \mid \text{Bag 2})P(\text{Bag 2})}{P(\text{Mango})}$$

- The probability of picking a fruit from bag 2 is the probability of getting tails on the coin flip:
$$P(\text{Bag 2}) = 0.4$$
- The probability of picking a mango from bag 2 is:
$$P(\text{Mango} \mid \text{Bag 2}) = \frac{4}{8} = 0.5$$
- The probability of picking a mango from bag 1 is:
$$P(\text{Mango} \mid \text{Bag 1}) = \frac{4}{6} = \frac{2}{3}$$
- The probability of picking a fruit from bag 1 is the probability of getting heads on the coin flip:
$$P(\text{Bag 1}) = 0.6$$

Calculating P(Mango)

The total probability of picking a mango is:

$$P(\text{Mango}) = P(\text{Mango} \mid \text{Bag 1})P(\text{Bag 1}) + P(\text{Mango} \mid \text{Bag 2})P(\text{Bag 2})$$

Substituting the values:

$$P(\text{Mango}) = \left(\frac{2}{3} \times 0.6\right) + (0.5 \times 0.4)$$

$$P(\text{Mango}) = 0.4 + 0.2 = 0.6$$

Applying Bayes' Theorem

Substituting the above values into Bayes' Theorem:

$$P(\text{Bag 2} \mid \text{Mango}) = \frac{P(\text{Mango} \mid \text{Bag 2})P(\text{Bag 2})}{P(\text{Mango})}$$

$$P(\text{Bag 2} \mid \text{Mango}) = \frac{0.5 \times 0.4}{0.6}$$

$$P(\text{Bag 2} \mid \text{Mango}) = \frac{0.2}{0.6} = \frac{1}{3}$$

Thus, the probability that the mango was picked from **bag 2** is:

$$\boxed{\frac{1}{3}}$$

Question 8

Given,

$$x_{t+1} = \mathbf{A}x_t + \omega, \quad \omega \sim \mathcal{N}(0, Q)$$
$$y_t = \mathbf{C}x_t + v, \quad v \sim \mathcal{N}(0, R)$$

where ω and v are i.i.d. gaussian noise variables and p(x$_0$) $= \mathcal{N}(\mu_0, \Sigma_0)$.

a. $p(x_0, x_1, \ldots, x_T)$

Since the state transition model is linear and Gaussian, the joint distribution of the states $x_0, x_1, \ldots, x_T$, by chain rule, can be written as:

$$p(x_0, x_1, \ldots, x_T) = p(x_0) \prod_{t=0}^{T-1} p(x_{t+1}|x_t)$$

Since $p(x_0) = \mathcal{N}(\mu_0, \Sigma_0)$, and the transition model is $x_{t+1} = \mathbf{A}x_t + \omega$, where $\omega \sim \mathcal{N}(0, Q)$, the conditional distribution $p(x_{t+1}|x_t)$ is Gaussian:

$$p(x_{t+1}|x_t) = \mathcal{N}(\mathbf{A}x_t, Q)$$

Thus, the joint distribution of $p(x_0, x_1, \ldots, x_T)$ is also a Gaussian,

$$p(x_0, x_1, \ldots, x_T) = \mathcal{N}(\mu_0, \Sigma_0) \prod_{t=0}^{T-1} \mathcal{N}(\mathbf{A}x_t, Q)$$

b. Given $p(x_t|y_1, \ldots, y_T) = \mathcal{N}(\mu_t, \Sigma_t)$

1. $p(x_{t+1}|y_1, \ldots, y_T)$

Since $x_{t+1} = \mathbf{A}x_t + \omega$, and $x_t|y_1, \ldots, y_T \sim \mathcal{N}(\mu_t, \Sigma_t)$, we can use the linear transformation property of Gaussian distributions. Thus,

$$p(x_{t+1}|y_1, \ldots, y_T) = \mathcal{N}(x_{t+1}|\mathbf{A}\mu_t, \mathbf{A}\Sigma_t\mathbf{A}^T + Q)$$

2. $p(x_{t+1}, y_{t+1}|y_1, \ldots, y_T)$

The joint distribution of x_{t+1} and y_{t+1} given the observations up to time T can be computed as follows:

$$p(x_{t+1}, y_{t+1}|y_1, \ldots, y_T) = p(x_{t+1}|y_1, \ldots, y_T)p(y_{t+1}|x_{t+1})$$

Since $p(x_{t+1}|y_1, \ldots, y_T) = \mathcal{N}(x_{t+1}|\mathbf{A}\mu_t, \mathbf{A}\Sigma_t\mathbf{A}^T + Q)$ and $y_{t+1} = \mathbf{C}x_{t+1} + v$ where $v \sim \mathcal{N}(0, R)$, we can write:

$$p(y_{t+1}|x_{t+1}) = \mathcal{N}(y_{t+1}|\mathbf{C}x_{t+1}, R)$$

Therefore, the joint distribution is:

$$p(x_{t+1}, y_{t+1}|y_1, \ldots, y_T) = \mathcal{N}(x_{t+1}|\mathbf{A}\mu_t, \mathbf{A}\Sigma_t\mathbf{A}^T + Q)\mathcal{N}(x_{t+1}|\mathbf{C}x_{t+1}, R)$$

3.$p(x_{t+1}|y_1, \ldots, y_{t+1})$

Given, $y_{t+1} = \hat{y}$, the conditional distribution $p(x_{t+1}|y_1, \ldots, y_{t+1})$ is obtained by updating the posterior based on the new observation. This is a Kalman filter update step, hence the new mean and covariance are:
First, the joint distribution $p(x_{t+1}, y_{t+1}|y_1, \ldots, y_T)$ is:

$$p(x_{t+1}, y_{t+1}|y_1, \ldots, y_T) = \mathcal{N}(\mathbf{A}\mu_t, \mathbf{A}\Sigma_t\mathbf{A}^T + Q)\mathcal{N}(\mathbf{C}x_{t+1}, R)$$

The conditional distribution is then:

$$p(x_{t+1}|y_1, \ldots, y_{t+1}) = \mathcal{N}(\mu_{t+1}, \Sigma_{t+1})$$

where

$$\mu_{t+1} = \mathbf{A}\mu_t + K(\hat{y} - \mathbf{C}\mathbf{A}\mu_t)$$
$$\Sigma_{t+1} = \mathbf{A}\Sigma_t\mathbf{A}^T + Q - K\mathbf{C}(\mathbf{A}\Sigma_t\mathbf{A}^T + Q)K^T$$

Here, K is the Kalman gain, given by:

$$K = (\mathbf{A}\Sigma_t\mathbf{A}^T + Q)\mathbf{C}^T(\mathbf{C}(\mathbf{A}\Sigma_t\mathbf{A}^T + Q)\mathbf{C}^T + R)^{-1}$$

Question 9

Part (a): Likelihood $p(y|x)$

The random variable y is defined as:

$$y = Ax + b + \omega,$$

where $\omega \sim \mathcal{N}(\omega|0, Q)$. Since ω is independent Gaussian noise, the conditional distribution $p(y|x)$ can be directly written as:

$$p(y|x) = \mathcal{N}(y|Ax + b, Q),$$

where:

- Mean: $\mu_y = Ax + b$,
- Covariance: $\Sigma_y = Q$.

Part (b): Marginal Distribution $p(y)$

The marginal distribution $p(y)$ is obtained by:

$$p(y) = \int p(y|x)p(x)dx,$$

where:

- $p(x) = \mathcal{N}(x|\mu_x, \Sigma_x)$,
- $p(y|x) = \mathcal{N}(y|Ax + b, Q)$.

The marginal $p(y)$ is Gaussian with mean and covariance:

1. **Mean:**

$$\mu_y = \mathbb{E}[y] = \mathbb{E}[Ax + b + \omega] = A\mathbb{E}[x] + b + \mathbb{E}[\omega] = A\mu_x + b.$$

2. **Covariance:**

$$\Sigma_y = \mathbb{E}[(y - \mu_y)(y - \mu_y)^T].$$

   Using independence of x and ω, we have:

$$\Sigma_y = A\Sigma_x A^T + Q.$$

Thus:

$$p(y) = \mathcal{N}(y|A\mu_x + b, A\Sigma_x A^T + Q).$$

Part (c): Measurement Mapping $z = Cy + v$

The random variable z is given by:

$$z = Cy + v,$$

where $v \sim \mathcal{N}(v|0, R)$.

(i) Likelihood $p(z|y)$

The conditional distribution $p(z|y)$ is:

$$p(z|y) = \mathcal{N}(z|Cy, R),$$

where:

- Mean: $\mu_z = Cy$,
- Covariance: $\Sigma_z = R$.

(ii) Marginal Distribution $p(z)$

The marginal distribution $p(z)$ is:

$$p(z) = \int p(z|y)p(y)dy.$$

Since $p(y)$ and $p(z|y)$ are Gaussian, $p(z)$ is also Gaussian. Its parameters are:

1. **Mean:**
$$\mu_z = C\mu_y = C(A\mu_x + b).$$
2. **Covariance:**
$$\Sigma_z = C\Sigma_y C^T + R,$$
where $\Sigma_y = A\Sigma_x A^T + Q$. Substituting:
$$\Sigma_z = C(A\Sigma_x A^T + Q)C^T + R.$$

Thus:
$$p(z) = \mathcal{N}(z|C(A\mu_x + b), C(A\Sigma_x A^T + Q)C^T + R).$$

Part (d): Posterior Distribution $p(x|\hat{y})$

To compute $p(x|\hat{y})$, we start with the joint Gaussian distribution $p(x, y)$.

(i) Joint Distribution $p(x, y)$

The random variables x and y are jointly Gaussian:

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix}\right),$$

where:

- $\Sigma_{xy} = \text{Cov}(x, y) = \text{Cov}(x, Ax + b + \omega) = A\Sigma_x,$
- $\Sigma_{yx} = \Sigma_{xy}^T = \Sigma_x A^T,$
- $\Sigma_y = A\Sigma_x A^T + Q.$

(ii) Posterior $p(x|\hat{y})$

The posterior $p(x|\hat{y})$ is Gaussian:

$$p(x|\hat{y}) = \mathcal{N}(x|\mu_{x|\hat{y}}, \Sigma_{x|\hat{y}}),$$

where:

1. **Mean:**
$$\mu_{x|\hat{y}} = \mu_x + \Sigma_{xy}\Sigma_y^{-1}(\hat{y} - \mu_y).$$
2. **Covariance:**
$$\Sigma_{x|\hat{y}} = \Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}.$$

Substituting $\Sigma_{xy} = A\Sigma_x$ and $\Sigma_y = A\Sigma_x A^T + Q$, we can compute the exact expressions for $\mu_{x|\hat{y}}$ and $\Sigma_{x|\hat{y}}$.

Question 10

```python
import numpy as np

N = 1000000

samples = np.random.randint(0, N+1, N)
samples = [i/N for i in samples]

A_N = np.mean(samples)

X = (A_N - 0.5) / (2 * np.sqrt(N))

print(f"Average A_N: {A_N}")
print(f"Standardized variable X: {X}")
```

Results:

$$\text{Average } A_N = 0.500331632631$$
$$\text{Standardized variable } X = 1.658163155000003 \times 10^{-7}$$

Conclusions:

- The sample average $A_N = 0.500331632631$ is remarkably close to the theoretical expected value of $\frac{1}{2}$, differing only by about 0.066%.
- The standardized variable X is very small ($\approx 1.66 \times 10^{-7}$), indicating that our sample mean is well within expected statistical fluctuations from the theoretical mean.
- This result supports the Law of Large Numbers, as with $N = 1,000,000$ samples, we achieve a very close approximation to the theoretical expected value.

- The small value of X also suggests that our random number generation is working properly, as it produces results consistent with theoretical expectations.

Here's a cleaner, more structured, and visually appealing version of your content with improved formatting and minor adjustments to improve clarity:

Question 11: Exploring the Penguin Dataset

Dataset Overview

Species

The dataset contains observations of three penguin species:

- **Adelie**
- **Chinstrap**
- **Gentoo**

Island

Penguins were observed across three islands:

- **Torgersen**
- **Biscoe**
- **Dream**

Physical Measurements

The dataset includes the following measurements:

- **Bill Length (mm):** Length of the penguin's bill.
- **Bill Depth (mm):** Depth of the penguin's bill.
- **Flipper Length (mm):** Length of the penguin's flipper.
- **Body Mass (g):** Mass of the penguin.

Sex

Penguins are categorized as either **Male** or **Female**.

Key Observations

Data Quality

- The dataset contains **344 rows and 7 columns**.
- Some missing values are present:
  - 2 rows have missing physical measurements.
  - 11 rows lack sex classifications.

Species Distribution

- **Adelie** penguins make up the largest portion of the dataset.
- **Gentoo** and **Chinstrap** penguins are also well-represented.

Measurement Ranges

- **Bill Length:** 33–60 mm
- **Bill Depth:** 13–21 mm
- **Flipper Length:** 170–230 mm
- **Body Mass:** 2700–6300 g

Sexual Dimorphism

- **Males** are generally larger than females in all measurements.
- This pattern holds across all three species.

Performing Exploratory Data Analysis (EDA)

Python Code for EDA

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load the dataset
url = "https://raw.githubusercontent.com/mwaskom/seaborn-data
   /master/penguins.csv"
df = pd.read_csv(url)

# Basic exploration
print("Dataset Shape:", df.shape)
print("Missing Values:\n", df.isnull().sum())

# Summary statistics
print("Summary Statistics by Species:")
print(df.groupby('species').describe())
```

Output:

- **Dataset Shape:** (344, 7)
- **Missing Values:**
  - Physical Measurements: 2 missing rows
  - Sex Classification: 11 missing rows

Correlation Analysis

```
# Numerical columns for correlation
numeric_cols = ['bill_length_mm', 'bill_depth_mm', '
    flipper_length_mm', 'body_mass_g']

# Correlation matrix
correlation_matrix = df[numeric_cols].corr()

# Heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm',
    center=0)
plt.title('Correlation Matrix of Penguin Measurements')
plt.tight_layout()
plt.show()
```

what is tokens in c programming language **Observations:**

- **Body Mass** has a strong positive correlation with **Flipper Length**.
- Weak correlations are observed between **Bill Length** and **Bill Depth**.

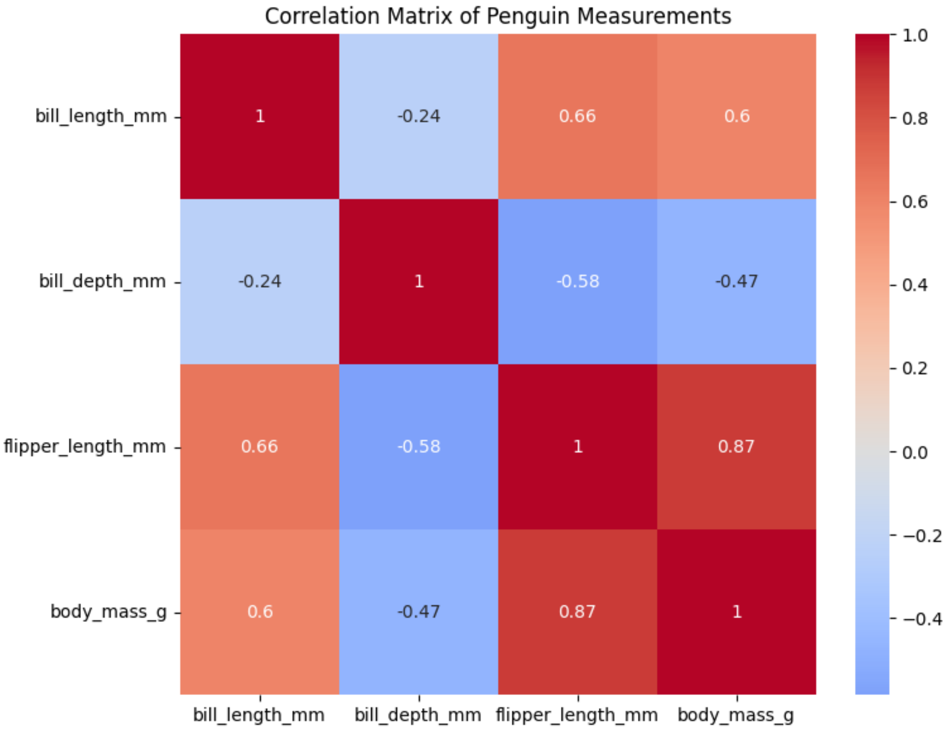


Figure 1: Correlation Matrix of Penguin Measurements

Distribution Analysis by Species and Sex

```python
# Boxplots for species and sex
plt.figure(figsize=(10, 7))
for i, column in enumerate(numeric_cols, 1):
    plt.subplot(2, 2, i)
    sns.boxplot(x='species', y=column, data=df, hue='sex')
    plt.title(f'{column} by Species and Sex')
plt.tight_layout()
plt.show()
```

Key Insights:

- **Gentoo** penguins tend to have the largest flipper lengths and body mass.
- **Adelie** penguins have smaller physical measurements compared to other species.
- Clear sexual dimorphism is observed, with males generally larger than females.

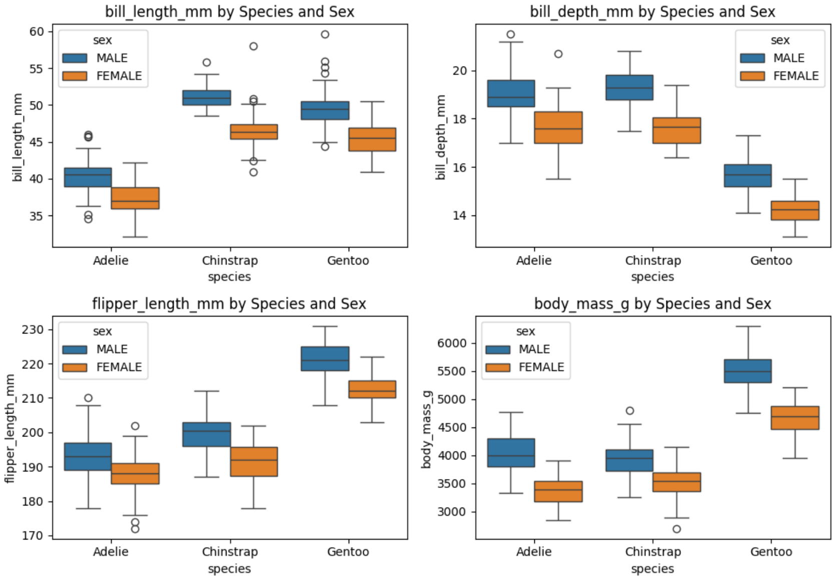


Figure 2: Measurement Distribution by Species and Sex

Pair Plots

```
# Pair plots
sns.pairplot(df, hue='species', diag_kind='kde')
plt.show()
```

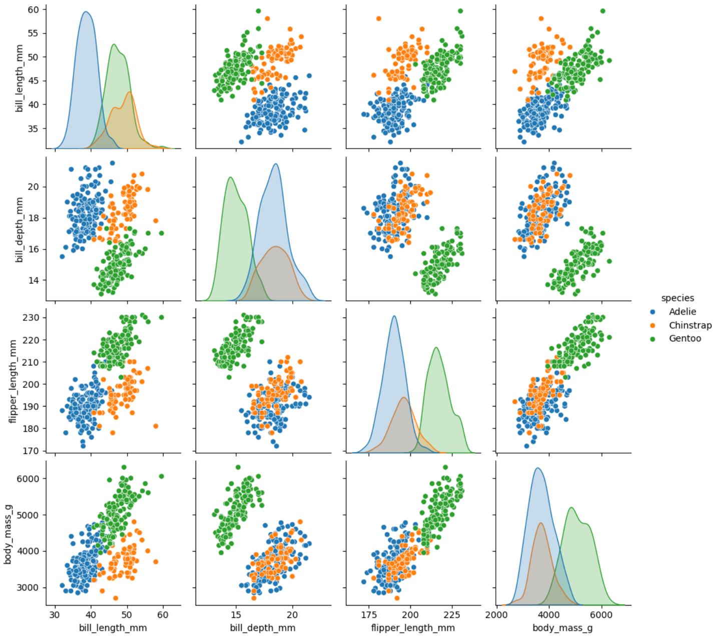


Figure 3: Pair Plot of Features of the Dataset

Clustering

```python
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans

def prepare_data(df):
    df = df.dropna()
    features = ['bill_length_mm', 'bill_depth_mm', 'flipper_length_mm', 'body_mass_g']
    X = df[features]
    scaler = StandardScaler()
    X_scaled = scaler.fit_transform(X)
    return X_scaled, X, features


def perform_clustering(X_scaled, X, features, n_clusters):
    kmeans = KMeans(n_clusters=n_clusters, random_state=42)
    clusters = kmeans.fit_predict(X_scaled)
    return clusters

X_scaled, X, features = prepare_data(df)
clusters = perform_clustering(X_scaled, X, features, n_clusters=3)

comparison_df = pd.DataFrame({
    'Actual_Species': df.dropna()['species'],
    'Predicted_Cluster': clusters
})

plt.figure(figsize=(8, 6))
confusion_matrix = pd.crosstab(
    comparison_df['Actual_Species'],
    comparison_df['Predicted_Cluster']
)

sns.heatmap(confusion_matrix, annot=True, fmt='d', cmap='YlGnBu')
plt.title('Cluster vs Actual Species Comparison')
plt.show()
```

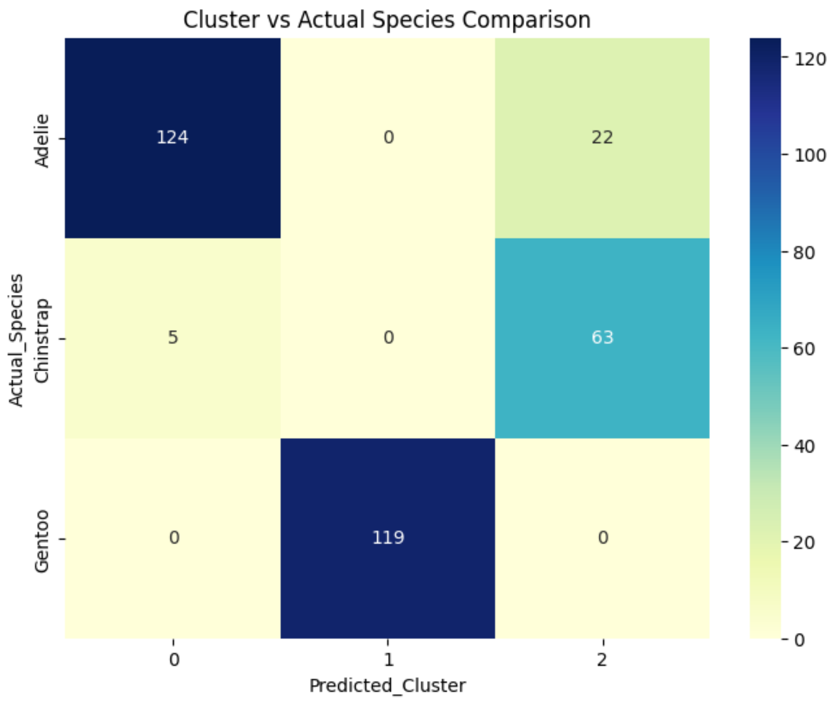


Figure 4: Confusion Matrix

Data Collection and Analysis: Comprehensive Analysis of Spotify Listening Data: April to December 2024

`Google Colab Notebook Link`

1. Introduction

This report details the findings from a detailed analysis of personal Spotify listening data, spanning April to December 2024. The primary objective was to uncover trends, patterns, and insights related to listening habits, artist preferences, song correlations, and session behavior. Data preprocessing, visualization, and statistical analysis techniques were employed to extract meaningful insights, with a strong emphasis on understanding the underlying user behavior.

2. Data Preprocessing

2.1. Data Collection

– Data was collected through an online service, Last.fm, which tracks the user's listening history on Spotify by integrating with the Spotify account.

– Last.fm gathers this information by either web scraping or through its API that provides structured listening history.

– The data used in this report was collected using the tool `lastfm-to-csv`, which enables efficient downloading of listening data in CSV format.

2.2. Source and Structure

– The dataset included four columns: `Artist`, `Album`, `Title`, and `Timestamp`.

– Data was fetched from the Last.fm/Spotify database, stored in CSV format.

– The CSV file was then uploaded to my github to fetch data into google colab `CSV File`

2.3. Timestamp Conversion

– Timestamps were converted to the **Indian Standard Time (IST)** zone for localized analysis.

– A new column, `Timestamp_IST`, was created for ease of interpretation.

– Timestamps were further decomposed into attributes such as **hour**, **day of the week**, **month**, and **date** to provide granular insights.

2.4. Data Cleaning

- Duplicates were removed, and missing values were checked. No significant issues were identified.

2.5. Derived Columns

- **Hour of Day**: Categorized into *Morning* (6–12), *Afternoon* (12–18), *Evening* (18–24), and *Late Night* (0–6).
- **Date** and **Day of Week**: Derived for time series analysis and weekly trends.
- **Session Attributes**: Calculated based on listening gaps and consecutive plays.

3. Time Series Analysis

3.1. Daily and Monthly Trends

- **Daily Listening Behavior**:
  - Activity peaked during **weekends**, particularly Saturdays, suggesting increased leisure time.
  - A sharp rise in listening frequency was noted during holidays and specific weeks in June.

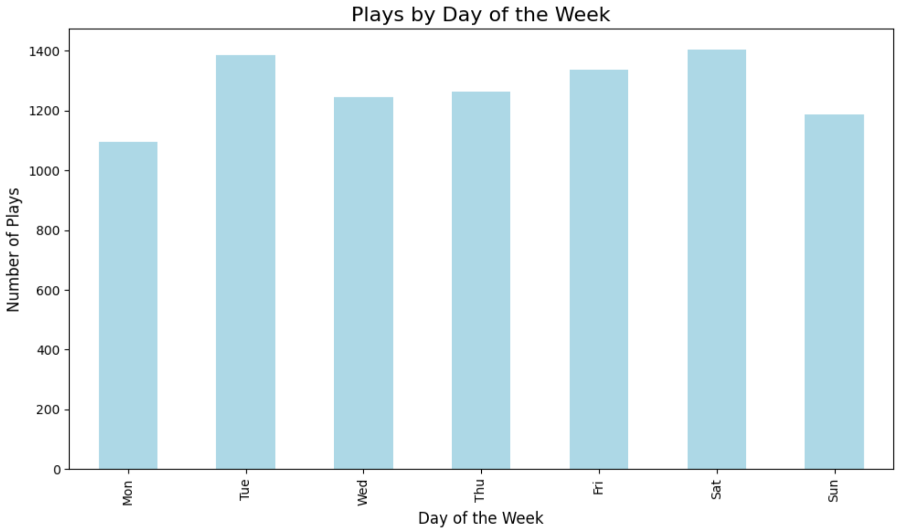


Figure 5: Plays by days of week

- **Monthly Listening Behavior**:
  - The highest listening activity was recorded in **June 2024**, with a total of 402 plays.
  - **November 2024** saw the least activity, attributed to academic or work commitments.

- **Reason**: The data can be explained by the following:
  - ∗ In June, I was working on a project in IIT Bhilai alone, hence the listening time suddenly spiked.
  - ∗ In October, my earphones broke, hence the sudden decline in the number of songs.
  - ∗ In November, my premium subscription ended, and the ads thereafter stopped me from listening to more music.
  - ∗ In December, I was back home, hence the listening time declined again.

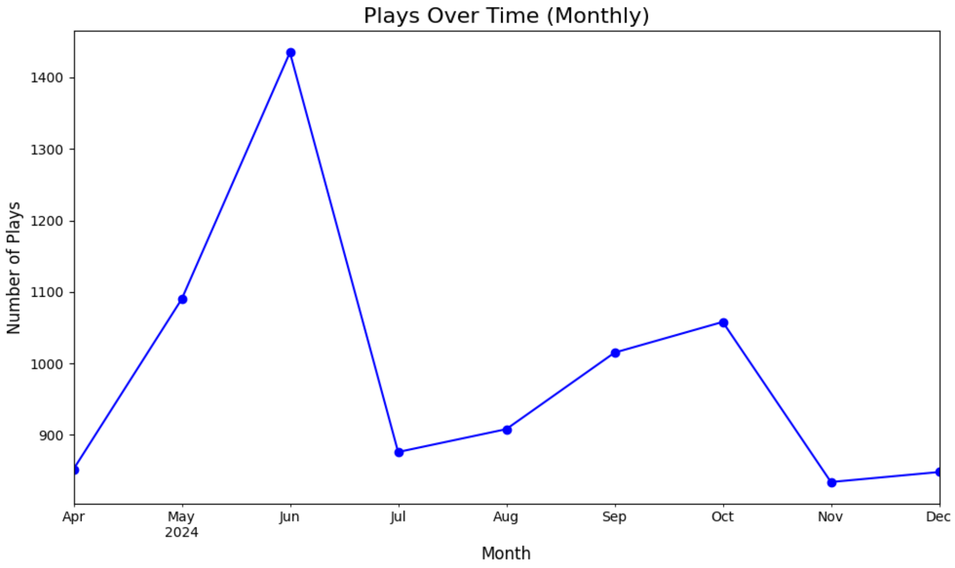


Figure 6: Plays Over Time(Monthly)

3.2. Hourly Patterns

- Listening activity was highest during the **evening hours (6 PM to 9 PM)**, correlating with leisure time after academic activites end at 5:30 PM.
- Late-night listening sessions were common on weekends, indicating a shift in routine and late sleeping times during the weekend.

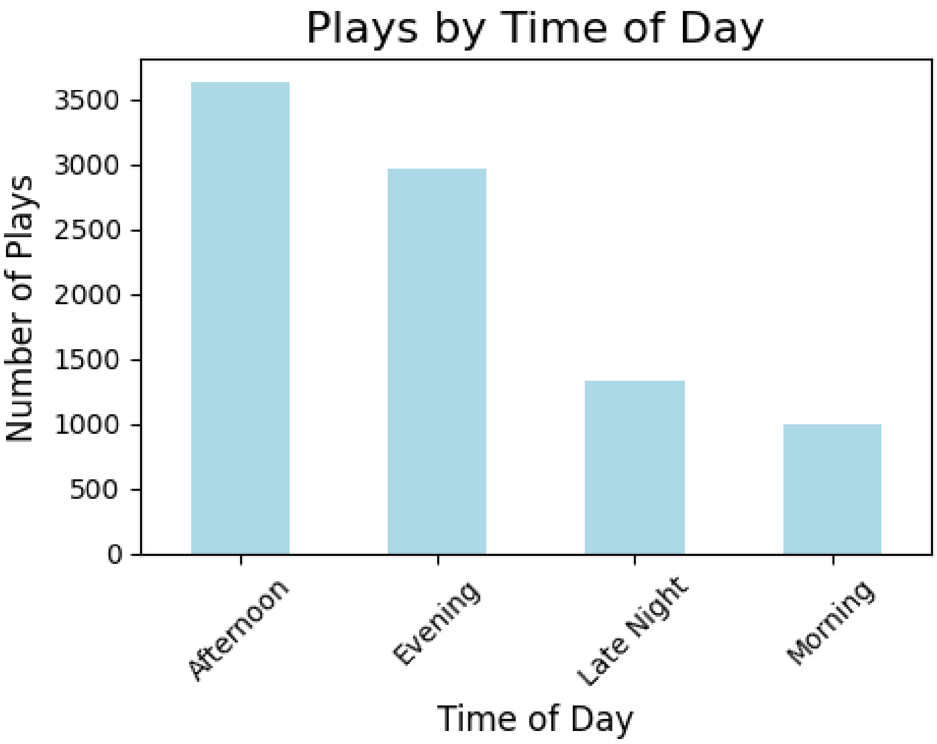


Figure 7: Plays by time of day

3.3. Peak Hours

- The top 3 peak hours were identified as 7 PM, 8 PM, and 9 PM.
- Visualized using **line plots**, with additional insights through radar charts showcasing hourly distributions.

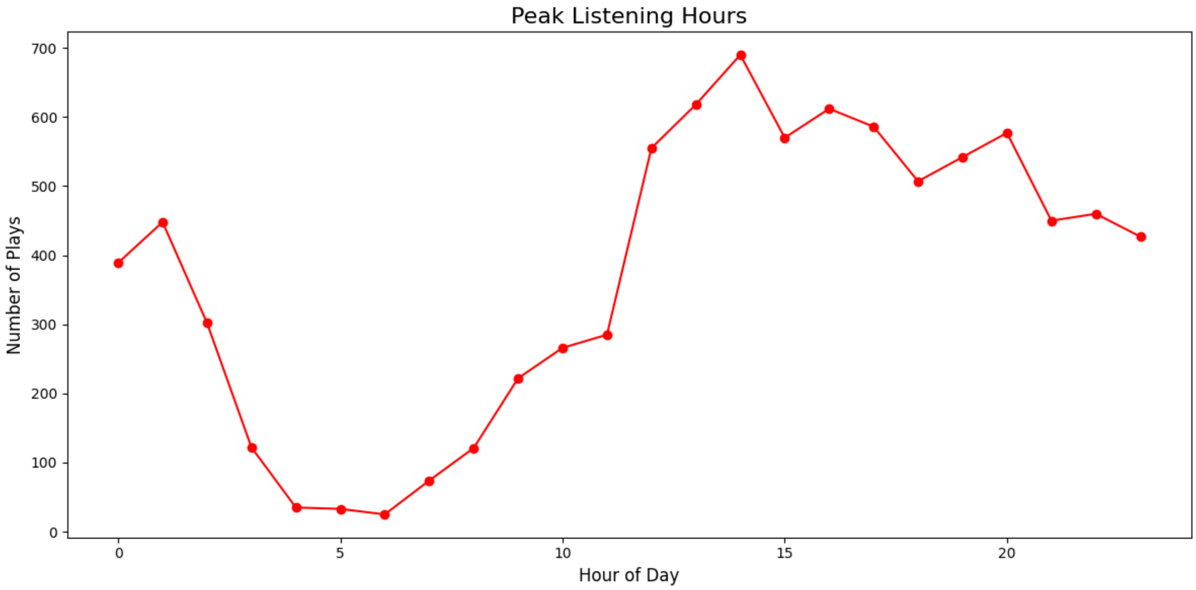


Figure 8: Peak Listening Hours

Insights:

- Listening habits align with typical patterns of relaxation, with evenings and weekends being the most active periods.
- Monthly peaks suggest specific events, seasons, or emotional states influence listening behavior.

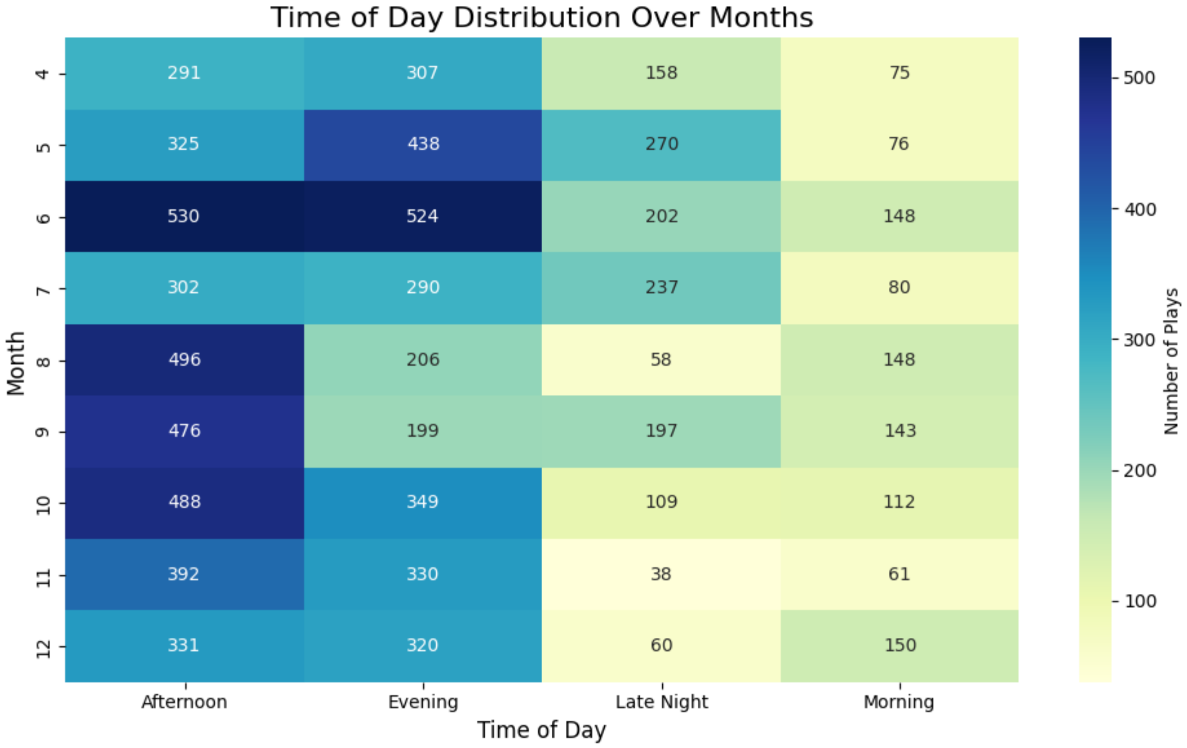


Figure 9: Time of Day Distribution Over Months

4. Favorite Artists, Songs, and Albums

4.1. Top Artists

- **Most Played Artists**:
  - **Anuv Jain**: 457 plays, with peak listening on June 4.
  - **Ariana Grande**: 399 plays, showcasing a broad appeal.
  - **Aditya Rikhari**: Consistent playtime across multiple months.
- Visualized using bar charts for the top 10 artists, revealing a preference for Indian and global indie artists.

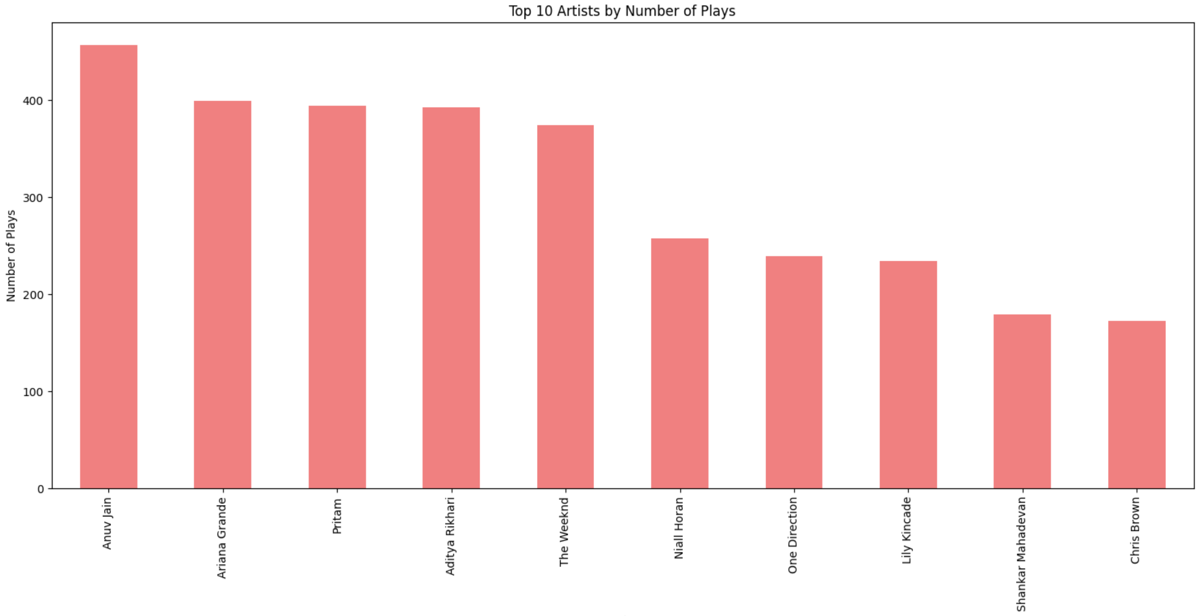


Figure 10: Top Artists

4.2. Top Songs

- **Consistent Favorites**:
  - *Faasle* by Aditya Rikhari: Appeared in 4 months, averaging 63.5 plays.
  - *Sweet n Low* by Lily Kincade: Appeared in 3 months, averaging 70.7 plays.
  - *Daylight* by David Kushner: Appeared in 2 months, averaging 47.5 plays.
  - *Co2* by Prateek Kuhad: Appeared in 2 months, averaging 35.5 plays.
  - *Breathless* by Shankar Mahadevan: Appeared in 2 months, averaging 34.5 plays.
  - *Samjho Na* by Aditya Rikhari: Appeared in 2 months, averaging 29.0 plays.

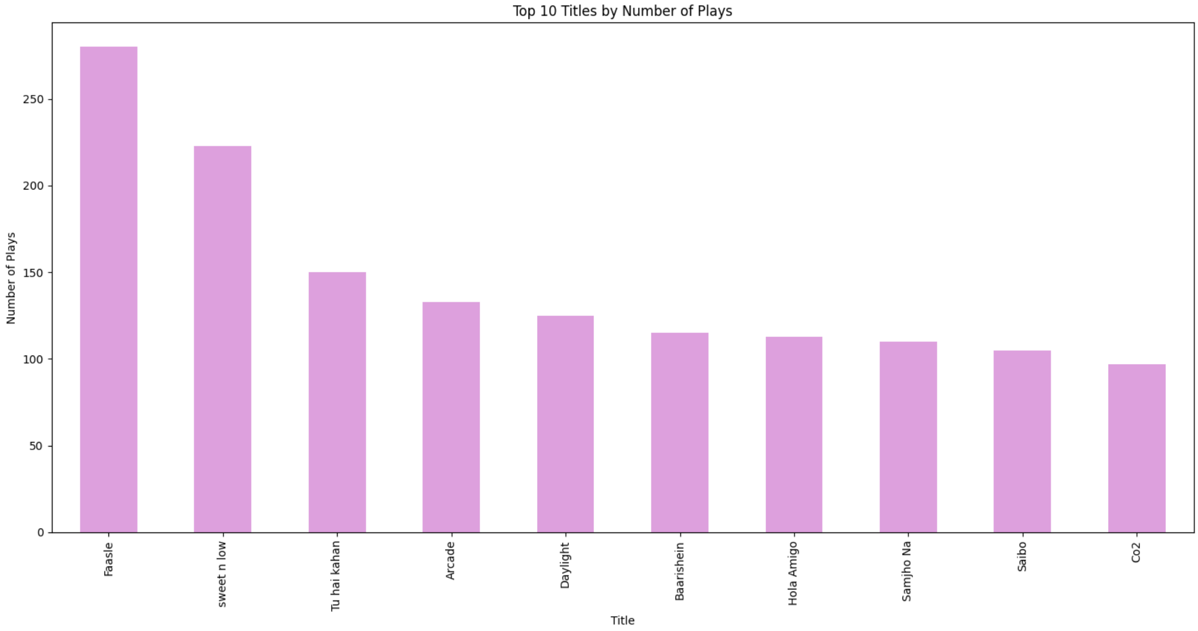


Figure 11: Top Songs

4.3. Top Albums

- Albums by Anuv Jain and Aditya Rikhari dominated the top 5, with significant daily plays.

4.4. Monthly Trends

- **2024-04**: 201 total plays, 5 unique artists, average plays per song: 40.2, top artist: Duncan Laurence.
- **2024-05**: 224 total plays, 5 unique artists, average plays per song: 44.8, top artist: Chris Brown.
- **2024-06**: 402 total plays, 4 unique artists, average plays per song: 80.4, top artist: Aditya Rikhari.
- **2024-07**: 99 total plays, 5 unique artists, average plays per song: 19.8, top artist: SZA.
- **2024-08**: 233 total plays, 5 unique artists, average plays per song: 46.6, top artist: Lily Kincade.
- **2024-09**: 269 total plays, 5 unique artists, average plays per song: 53.8, top artist: The Local Train.
- **2024-10**: 288 total plays, 4 unique artists, average plays per song: 57.6, top artist: Shankar Mahadevan.
- **2024-11**: 91 total plays, 4 unique artists, average plays per song: 18.2, top artist: Aditya Rikhari.
- **2024-12**: 74 total plays, 5 unique artists, average plays per song: 14.8, top artist: Lady Gaga.

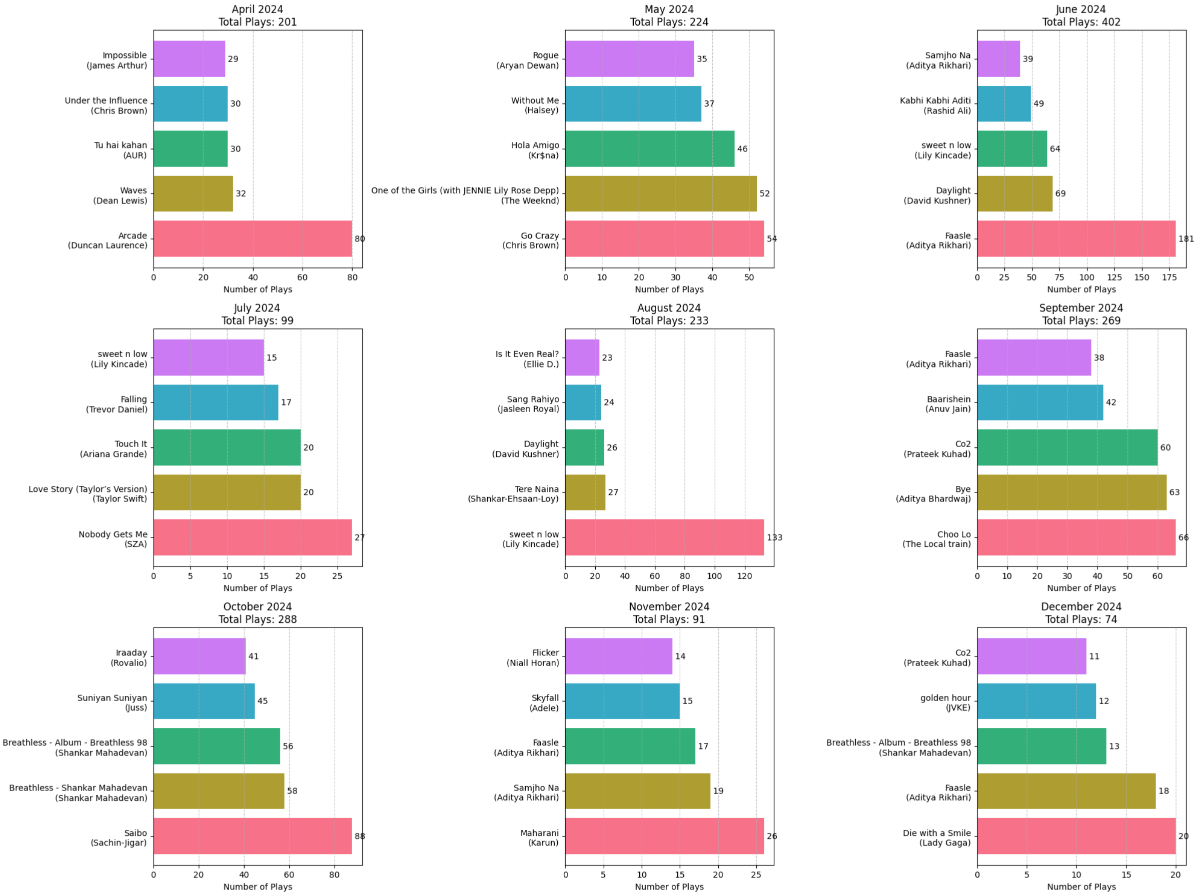


Figure 12: Monthly top songs

Insights:

- Repeated plays of specific songs and artists highlight emotional or thematic connections.
- A preference for a mix of Indian indie, Western pop, and alternative genres is evident.

5. Session Analysis

5.1. Definition of a Session

- **Criteria**:
  - ∗ A session was defined by a gap of fewer than 20 minutes between consecutive plays.
  - ∗ Minimum session size: 7 songs.

5.2. Metrics

- **Average Session Duration**: 62 minutes.

- **Average Songs per Session**: 15.35.
- **Most Active Period**: Afternoon and evening sessions on Tuesdays.

5.2.1. Session Analysis Summary

- **Total number of sessions**: 384
- **Average session duration**: 62.02 minutes
- **Average songs per session**: 15.35
- **Most common session start time**: Afternoon
- **Most active day**: Tuesday

5.3. Session Trends

- Visualized using boxplots and scatter plots.
- High artist diversity during longer sessions indicates exploratory listening behavior.

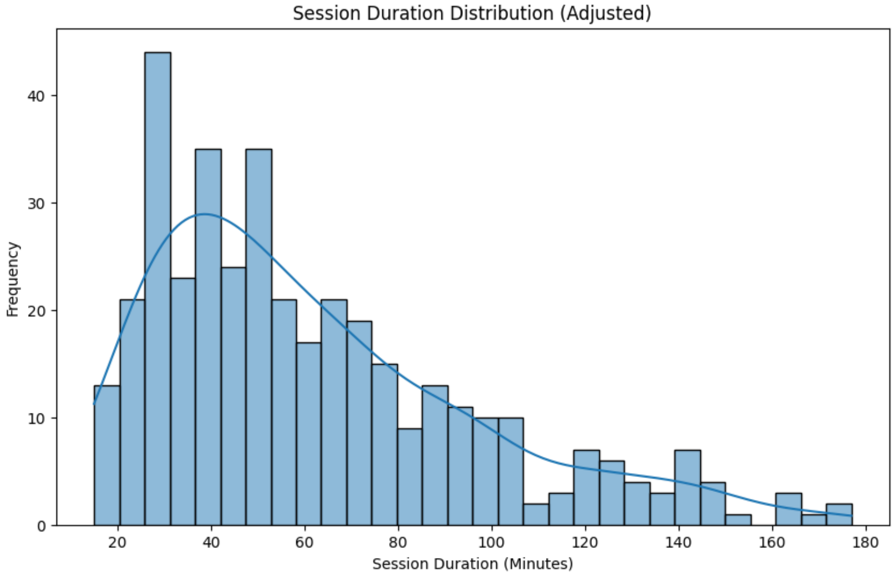


Figure 13: Session Duration Distribution

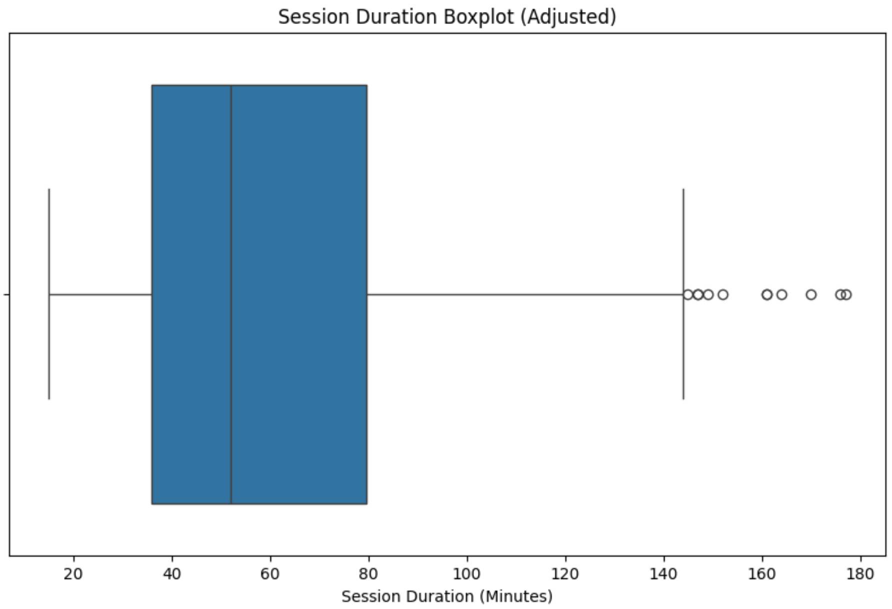


Figure 14: Session Box Plot

5.3.1. Engagement Metrics

- **Average songs per minute**: 0.26

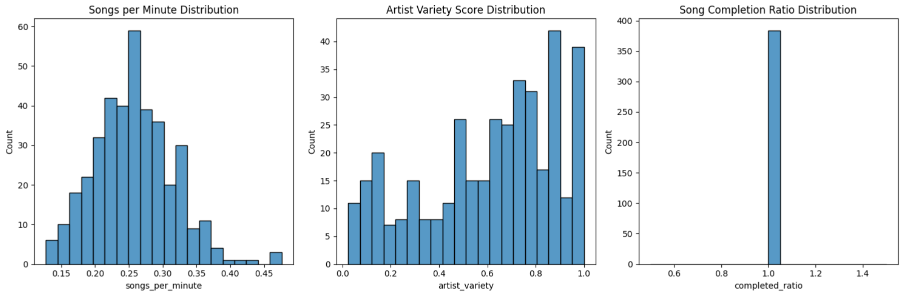


Figure 15: Engagement Metrics

Insights:

- Consistent afternoon sessions suggest music as a productivity or relaxation aid.
- Longer evening sessions align with leisure or social listening contexts.
- The high average number of songs per session and consistent start times (Afternoon, Tuesday) demonstrate that the listener engages in regular, extended listening sessions, often in structured time periods such as afternoons or evenings.

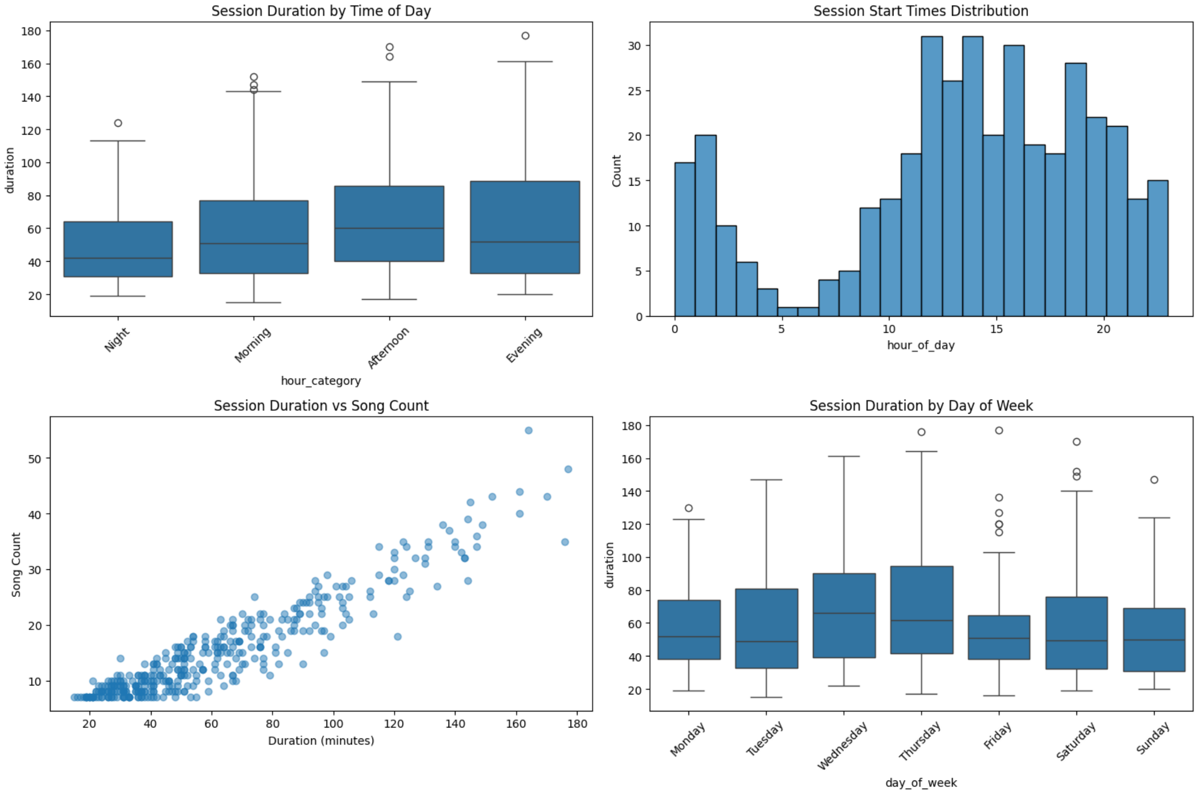


Figure 16: Enter Caption

6. Correlation Analysis

6.1. Artist Correlations

- **Strongest Positive Correlation**:
  - Anuv Jain and Anuv Jain (1.000): Perfect self-correlation.
- **Strongest Negative Correlation**:
  - Ariana Grande and Anuv Jain (-0.119): Contrasting listening patterns across different genres and moods.
- **Most Consistent Artist**:
  - Pritam, active on 124 days.
- **Highest Daily Play Intensity**:
  - Lily Kincade, with an average of 7.31 plays per active day.
- **Notable Patterns**:
  - Aditya Rikhari and Anuv Jain: 0.610
  - Ariana Grande and One Direction: 0.380

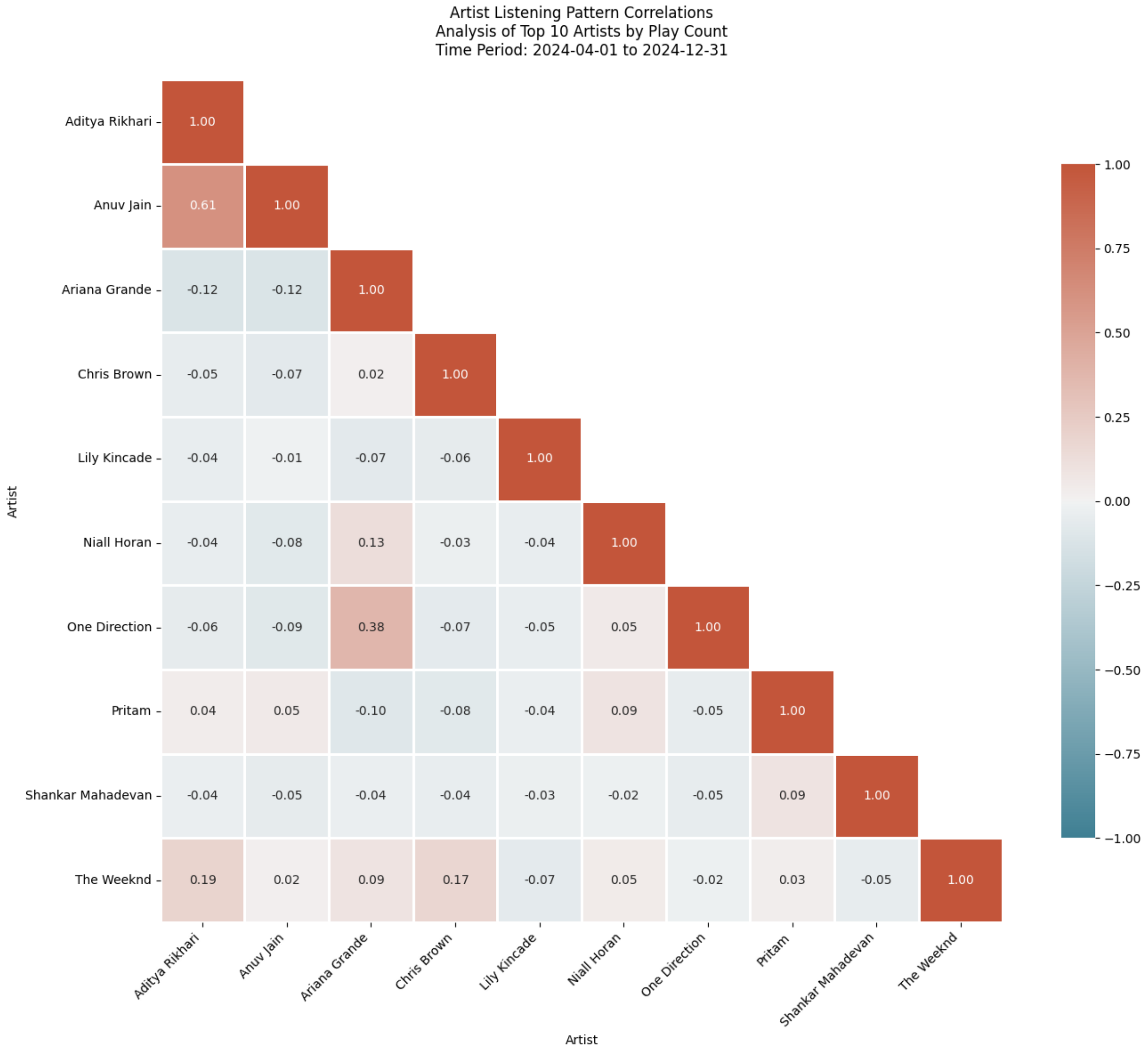


Figure 17: Artists Correlation

6.2. Song Correlations

- **Strongest Positive Correlation**:
  - ’Baarishein’ by Anuv Jain with itself (1.000).
- **Strongest Negative Correlation**:
  - ’Arcade’ by Duncan Laurence and ’Baarishein’ by Anuv Jain (-0.128).
- **Most Consistently Played Song**:
  - ’Faasle’ by Aditya Rikhari, active on 75 days.
- **Most Intensely Played Song**:
  - ’sweet n low’ by Lily Kincade, with an average of 7.19 plays per active day.
- **Notable Patterns**:
  - ’Baarishein’ (Anuv Jain) and ’Samjho Na’ (Aditya Rikhari): 0.693

- ∗ ’Samjho Na’ (Aditya Rikhari) and ’Baarishein’ (Anuv Jain): 0.693

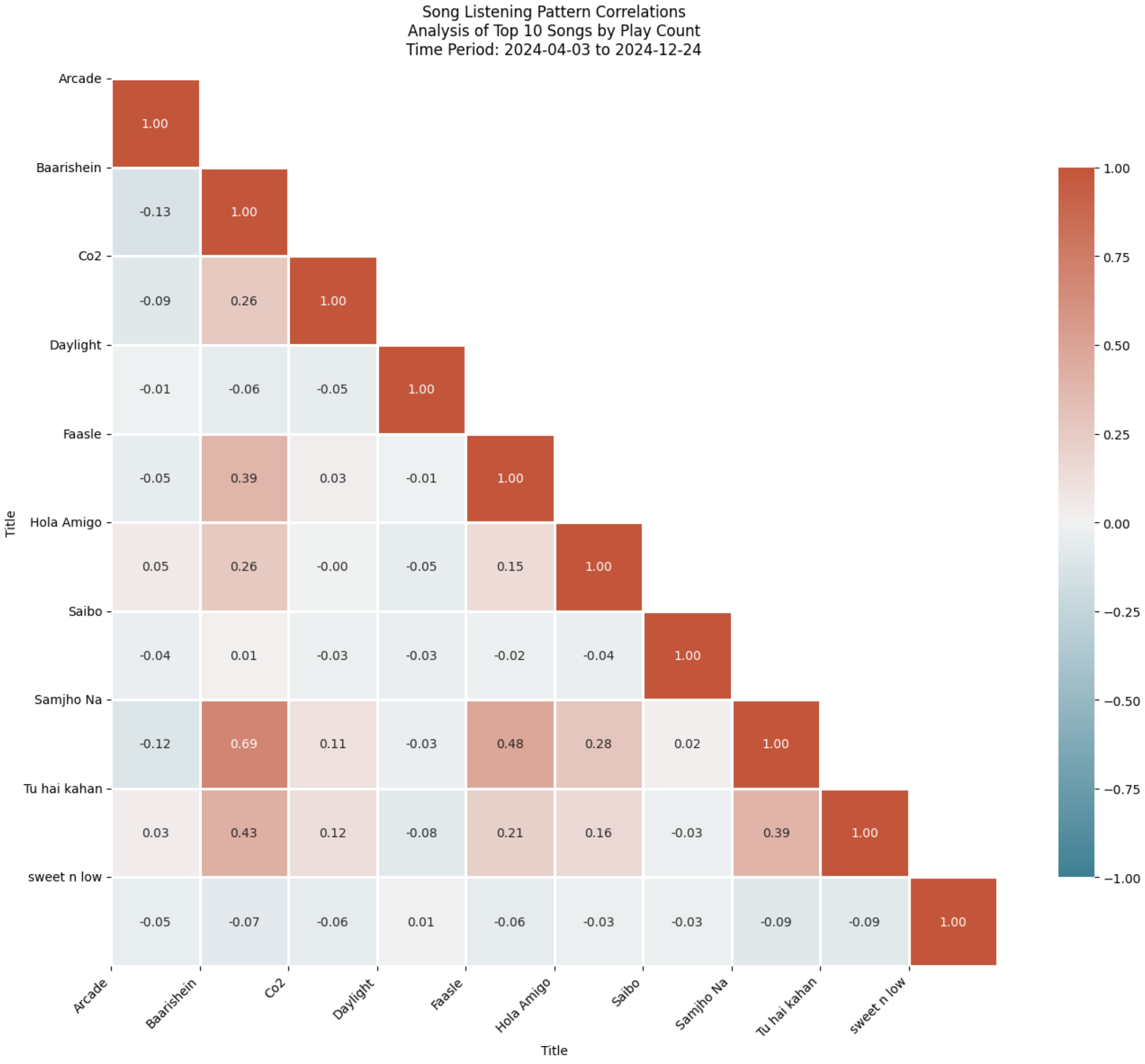
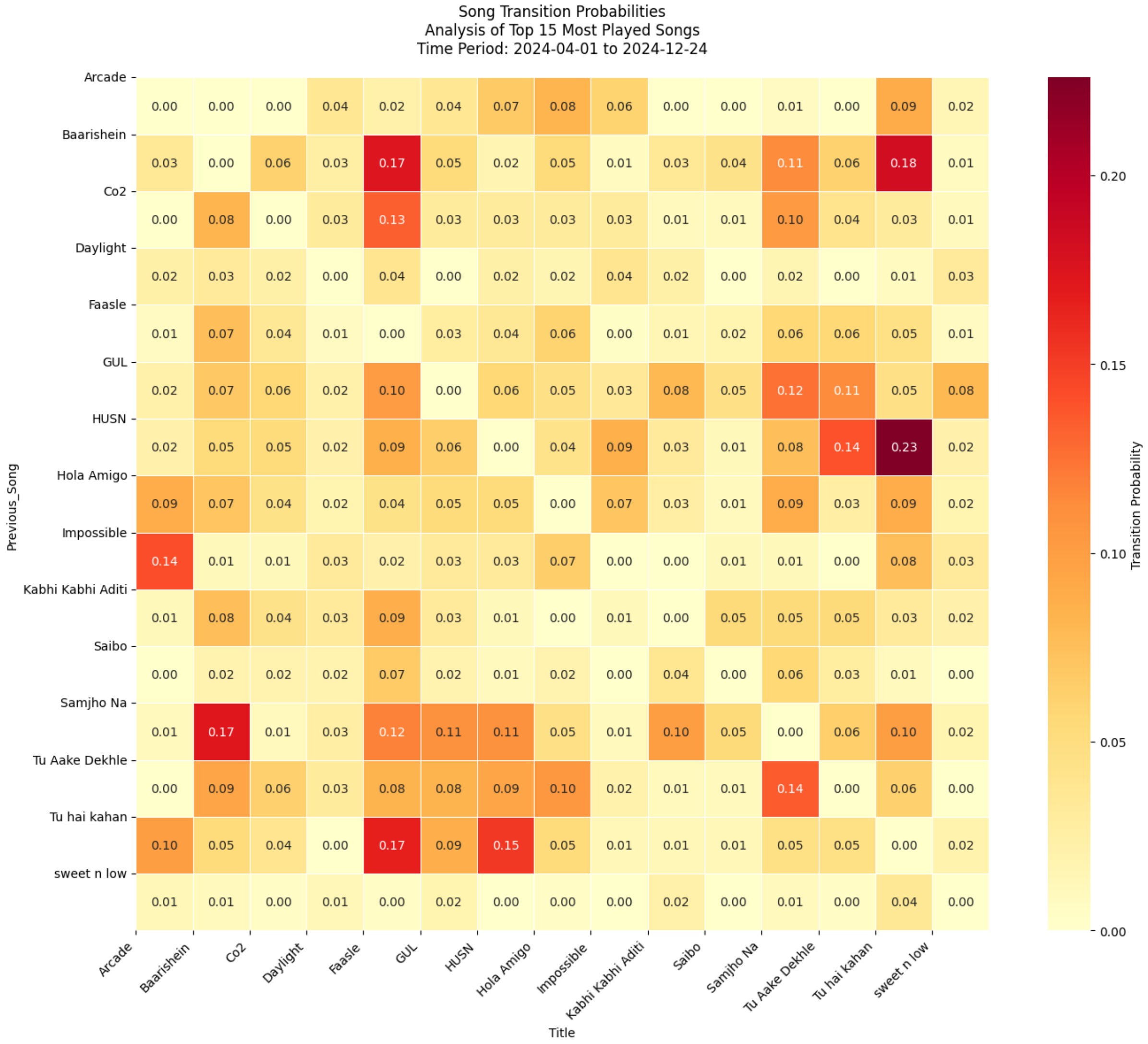


Figure 18: Songs Correlation

6.3. Song Transition Analysis

- **Strongest Song Transitions**:
  - ∗ *HUSN* by Anuv Jain → *Tu hai kahan* by AUR (22.58%).
  - ∗ *Baarishein* by Anuv Jain → *Tu hai kahan* by AUR (18.26%).
  - ∗ *Baarishein* by Anuv Jain → *Faasle* by Aditya Rikhari (17.39%).
- **Transition Timing**:
  - ∗ Average time between songs: 200.7 minutes.
  - ∗ Median time between songs: 6.0 minutes.
- **Artist Transition Patterns**:
  - ∗ Same artist transitions: 3.1%.

- ∗ Different artist transitions: 96.9%.



	Arcade	Baarishein	Co2	Daylight	Faasle	GUL	HUSN	Hola Amigo	Impossible	Kabhi Kabhi Aditi	Saibo	Samjho Na	Tu Aake Dekhle	Tu hai kahan	sweet n low
Arcade	0.00	0.00	0.00	0.04	0.02	0.04	0.07	0.08	0.06	0.00	0.00	0.01	0.00	0.09	0.02
Baarishein	0.03	0.00	0.06	0.03	0.17	0.05	0.02	0.05	0.01	0.03	0.04	0.11	0.06	0.18	0.01
Co2	0.00	0.08	0.00	0.03	0.13	0.03	0.03	0.03	0.03	0.01	0.01	0.10	0.04	0.03	0.01
Daylight	0.02	0.03	0.02	0.00	0.04	0.00	0.02	0.02	0.04	0.02	0.00	0.02	0.00	0.01	0.03
Faasle	0.01	0.07	0.04	0.01	0.00	0.03	0.04	0.06	0.00	0.01	0.02	0.06	0.06	0.05	0.01
GUL	0.02	0.07	0.06	0.02	0.10	0.00	0.06	0.05	0.03	0.08	0.05	0.12	0.11	0.05	0.08
HUSN	0.02	0.05	0.05	0.02	0.09	0.06	0.00	0.04	0.09	0.03	0.01	0.08	0.14	0.23	0.02
Hola Amigo	0.09	0.07	0.04	0.02	0.04	0.05	0.05	0.00	0.07	0.03	0.01	0.09	0.03	0.09	0.02
Impossible	0.14	0.01	0.01	0.03	0.02	0.03	0.03	0.07	0.00	0.00	0.01	0.01	0.00	0.08	0.03
Kabhi Kabhi Aditi	0.01	0.08	0.04	0.03	0.09	0.03	0.01	0.00	0.01	0.00	0.05	0.05	0.05	0.03	0.02
Saibo	0.00	0.02	0.02	0.02	0.07	0.02	0.01	0.02	0.00	0.04	0.00	0.06	0.03	0.01	0.00
Samjho Na	0.01	0.17	0.01	0.03	0.12	0.11	0.11	0.05	0.01	0.10	0.05	0.00	0.06	0.10	0.02
Tu Aake Dekhle	0.00	0.09	0.06	0.03	0.08	0.08	0.09	0.10	0.02	0.01	0.01	0.14	0.00	0.06	0.00
Tu hai kahan	0.10	0.05	0.04	0.00	0.17	0.09	0.15	0.05	0.01	0.01	0.01	0.05	0.05	0.00	0.02
sweet n low	0.01	0.01	0.00	0.01	0.00	0.02	0.00	0.00	0.00	0.02	0.00	0.01	0.00	0.04	0.00

Figure 19: Song Transition Correlation

Insights:

– Strong transitions between emotionally resonant songs suggest mood-based playlists.

– Contrasting artist correlations point to varied listening contexts (e.g., focused vs. relaxing).

– High consistency of Pritam and Lily Kincade shows a preference for specific genres and mood states.

7. Discovery and Engagement Patterns

7.1. Discovery Rates

- A sharp rise in new artist discoveries during June, suggesting active exploration during this period.

7.1.1. Discovery Summary and Insights

- The listener actively explored new artists during June, with a significant increase in discoveries observed.
- This indicates a potential period of heightened interest in diverse music or genre exploration.

7.2. Engagement Metrics

- Repeat listens accounted for **87.6%** of all plays, indicating deep engagement with favorite tracks.
- The consistency score for overall listening was 0.31, highlighting regular habits.

7.2.1. Engagement Summary and Insights

- The high percentage of repeat listens (87.6%) reflects a strong preference for familiar songs, showing that the listener has a set of favorites that they consistently return to.
- The consistency score of 0.31 further confirms regular listening patterns, indicating moderate to high engagement over time.

7.3. Artist Loyalty

- **Metrics**:
  - Lily Kincade had the highest loyalty score (2.44), attributed to consistent daily plays.

7.3.1. Artist Loyalty Summary and Insights

- **Lily Kincade** leads with the highest loyalty score of 2.44, attributed to her frequent plays across a consistent period. This indicates the listener's deep connection to her music.
- Other artists like **Shankar Mahadevan**, **Niall Horan**, and **Juss** also exhibit strong loyalty, although at slightly lower levels compared to Lily Kincade.

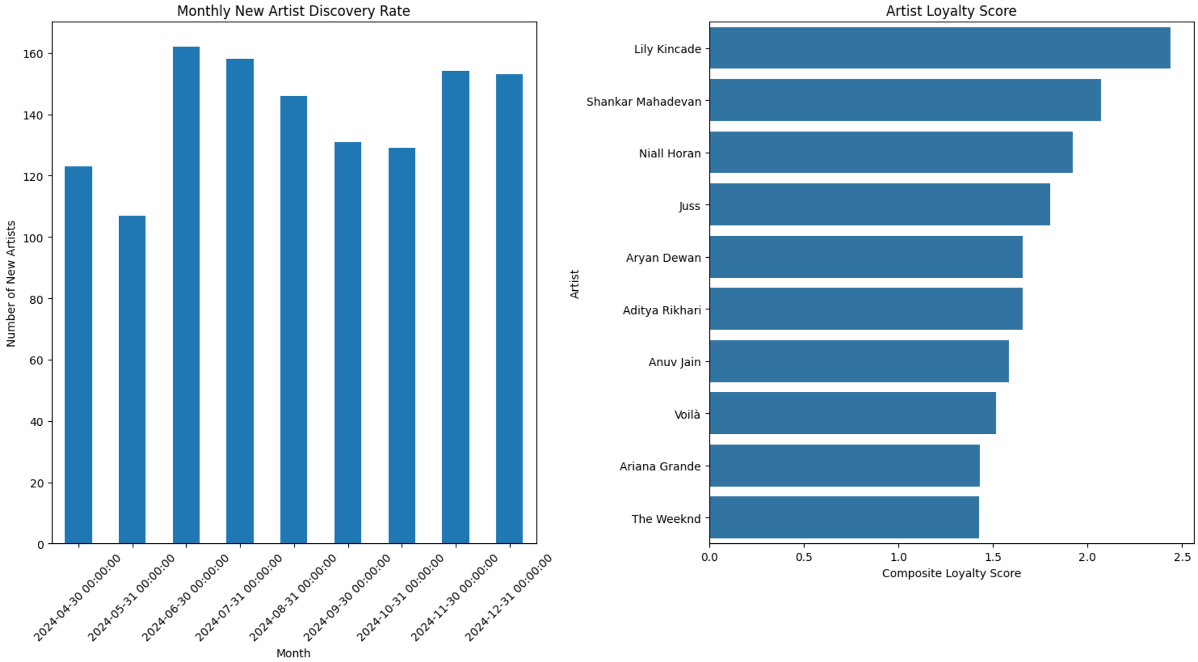


Figure 20: Artist Discovery Rate and Loyalty

7.4. Artist Trending Insights

- **Anuv Jain**: 457 total plays with an average of 1.66 plays per day. Peak on June 4.
- **Ariana Grande**: 399 total plays, averaging 1.45 plays per day. Peak on July 9.
- **Pritam**: 394 total plays, averaging 1.43 plays per day. Peak on September 24.
- **Aditya Rikhari**: 393 total plays, averaging 1.43 plays per day. Peak on June 3.
- **The Weeknd**: 374 total plays, averaging 1.36 plays per day. Peak on May 16.

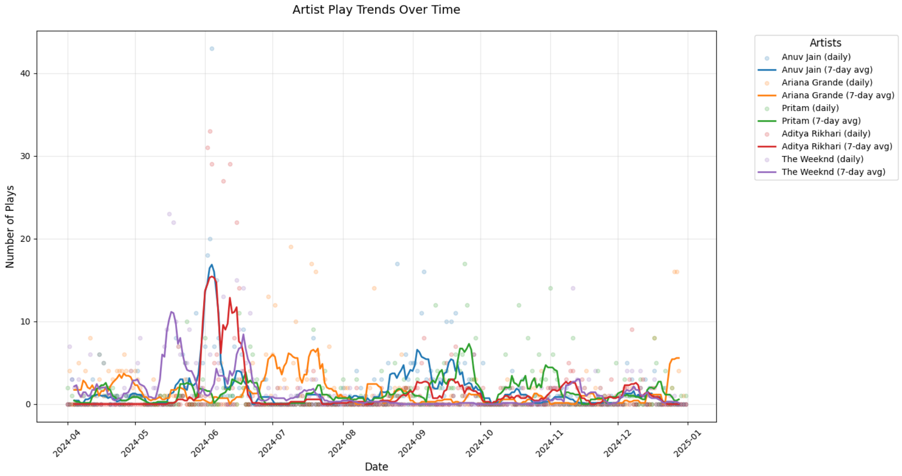


Figure 21: Artists play trend over time

7.5. Title Trending Insights

- *Faasle* by Aditya Rikhari: 280 plays with an average of 1.05 plays per day. Peak on June 3.
- *Sweet n Low* by Lily Kincade: 223 plays with an average of 0.84 plays per day. Peak on August 9.
- *Tu Hai Kahan*: 150 plays with an average of 0.56 plays per day. Peak on May 2.
- *Arcade*: 133 plays with an average of 0.50 plays per day. Peak on April 7.
- *Daylight*: 125 plays with an average of 0.47 plays per day. Peak on June 22.

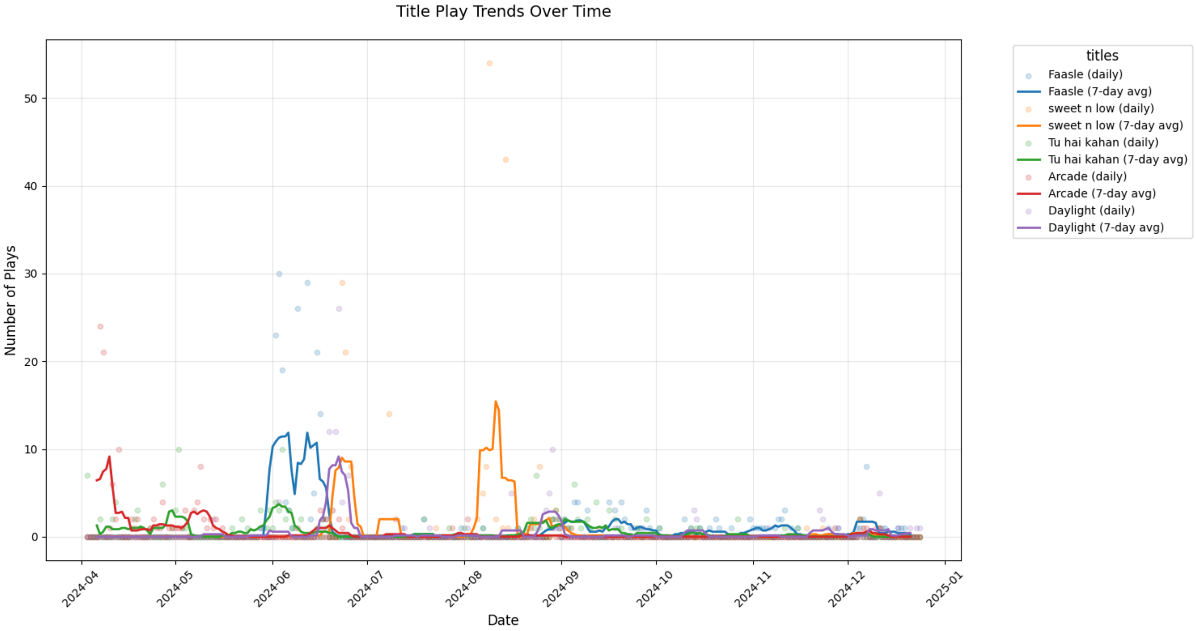


Figure 22: Title play trend over time

Insights:

- Discovery peaks align with changes in mood, seasons, or external influences.
- High repeat rates emphasize a preference for emotional familiarity.
- High loyalty scores for select artists indicate strong emotional connections.

8. Clustering Analysis

8.1. Hierarchical Clustering

- Artists were clustered based on hourly, daily, and monthly listening distributions, using hierarchical clustering with Ward's method.
- Distinct clusters emerged, separating indie, pop, and experimental artists, as well as artists with varying fanbase sizes and listening patterns.
- The hierarchical dendrogram provides insights into the relationships and similarity of artist listening patterns.

Cluster Analysis Results:

The clustering analysis identified the following groups:

Cluster	**Size**	**Artists (Top 5)**	**Percentage of Total Artists**
1	2	Niall Horan, One Direction	8%
2	1	David Kushner	4%
3	2	Rashid Ali, Sachin-Jigar	8%
4	10	Ariana Grande, Chris Brown, Duncan Laurence, H.E.R., Khalid	40%
5	10	AUR, Aditya Rikhari, Anuv Jain, King, Lily King	40%

Table 1: Clustering Results and Artist Distribution

Insights:

- The clustering highlights the diversity in artist listening patterns, with larger clusters corresponding to well-known pop and mainstream artists, while smaller clusters feature indie and experimental genres.
- The larger clusters (Clusters 4 and 5) represent a significant portion (80%) of the artists, indicating strong groupings based on genre or popularity.
- Smaller clusters (Clusters 1, 2, and 3) suggest more niche or unique listening patterns, likely reflecting distinct fanbase preferences and music styles.

Dendrogram Visualizations:

- The complete hierarchical dendrogram and truncated dendrogram provide a clear visual representation of the clusters. The full dendrogram showcases the distance between clusters, while the truncated version highlights the most distinct clusters.

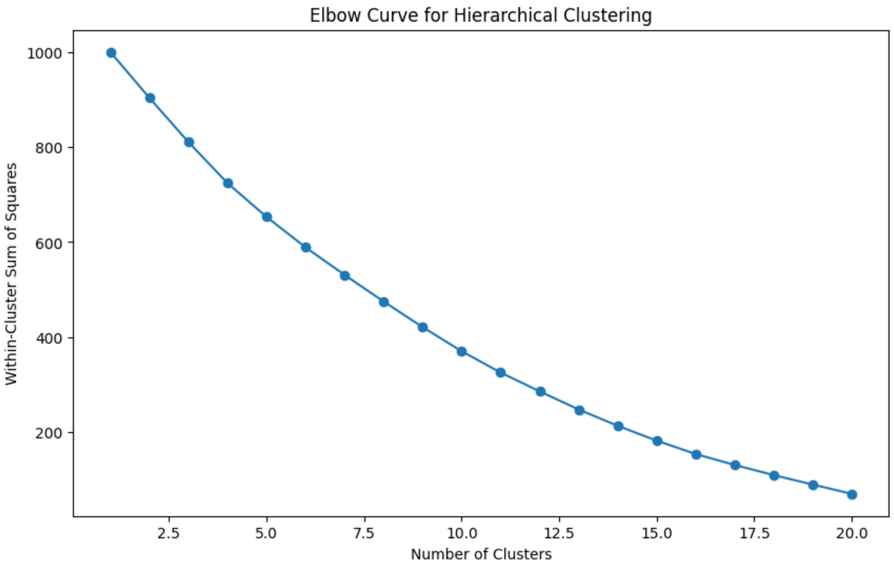


Figure 23: Cluster Elbow Curve

- An elbow curve analysis was performed to determine the optimal number of clusters. The elbow suggests a well-defined break in the data around 5 clusters, which was confirmed by further cluster analysis.

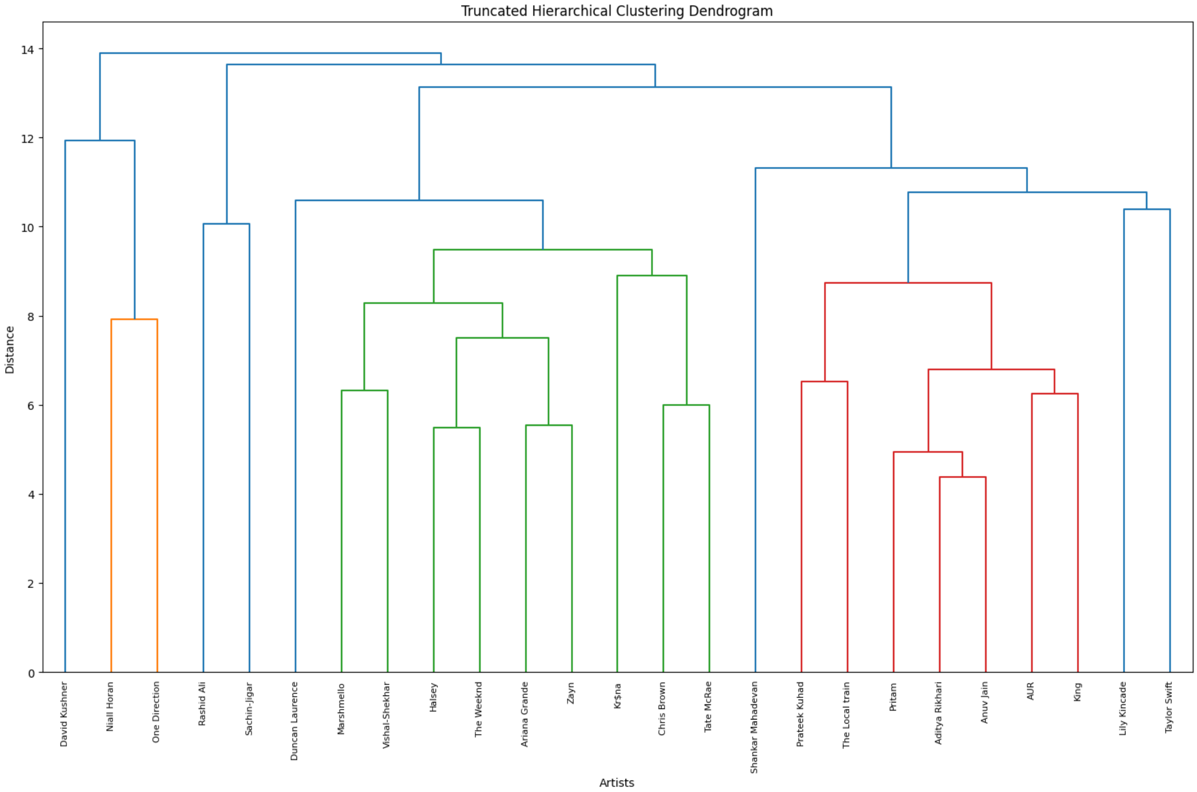


Figure 24: Cluters Dendogram

9. Conclusion

This analysis of Spotify listening data reveals intricate patterns in music consumption, driven by emotional, temporal, and contextual factors. Insights into artist loyalty, session behavior, and transition dynamics offer a deeper understanding of personal listening habits. Future studies could incorporate genre-specific analysis and predictive models to further enhance understanding.