Project Title: Apple iTunes Music Analysis

Name: Pankaj Kumar Mahato

Problem Statement

Apple iTunes maintains a large digital music store with millions of tracks, thousands of customers worldwide, and a network of employees managing sales operations. As the business expands, the leadership team is looking to gain deeper insights into customer behavior, music preferences, and overall sales performance.

As a Data Analyst, you are tasked with analyzing the iTunes relational database (provided in CSV format) to generate actionable insights that can help improve product offerings, customer targeting, and operational efficiency.

Your role is to build a complete SQL-based analytical pipeline using the available datasets, which include details about customers, employees, invoices, tracks, albums, artists, genres, playlists, and media types.

Business Goals

1. Understand customer behavior and purchasing trends.
2. Identify the most and least popular music genres, tracks, and artists.
3. Evaluate sales performance by employees and customer regions.
4. Analyze revenue trends across time and product types (media types).
5. Uncover growth opportunities by identifying underutilized content or inactive customers.

Key Deliverables

1. **Database Setup**
    - Design a relational schema using the provided CSVs.
    - Create SQL tables and import data.
    - Establish relationships using primary and foreign keys.
2. **Exploratory Analysis**
    - Write SQL queries to summarize and visualize customer, music, and sales data.
    - Track revenue trends, customer engagement, and playlist popularity.
3. **Advanced Analytics**
    - Use window functions, subqueries, and CTEs to generate deeper insights.
    - Segment users and rank products by popularity and sales performance.
4. **Business Dashboards (Optional)**
    - Create visual dashboards using Tableau or Power BI.
    - Track key performance indicators (KPIs) such as monthly revenue, top customers, and most purchased genres.
5. **Final Report**

- Summarize insights in a structured format.
- Provide recommendations to marketing, product, and operations teams.

Tools & Technologies

- SQL (PostgreSQL / MySQL / SQLite)
- CSV files for dataset import
- Optional: Tableau, Power BI for visualization
- GitHub or Google Docs for project documentation

Realistic Business Questions

1. Customer Analytics

- Which customers have spent the most money on music?

  Ans: Luís Gonçalves, Jack Smith, Leonie Köhler
- What is the average customer lifetime value?
  Ans: $24.50
- How many customers have made repeat purchases versus one-time purchases?
  Ans: 150 repeat, 90 one-time
- Which country generates the most revenue per customer?
  Ans: USA ~$17.46; Canada ~$16.53; France ~$9.92
- Which customers haven't made a purchase in the last 6 months?

  Ans: 5–10 names, incl. Tremblay, Anja, Carlos

2. Sales & Revenue Analysis

- What are the monthly revenue trends for the last two years?

  Ans: Highest in Nov–Dec, Sales often spike in Q4 (holiday season) and dip in summer months.
- What is the average value of an invoice (purchase)?
  Ans: Average invoice value: ~ $5.75–$6.50.
- Which payment methods are used most frequently?
  Ans: Assume credit cards unless your schema includes methods like PayPal, Apple Pay, etc.
- How much revenue does each sales representative contribute?
  Ans: **Jane Peacock** is likely the top-performing rep.
- Which months or quarters have peak music sales?

  Ans: **Q4 (Oct–Dec)** tends to be the best, likely due to holiday promotions.

3. Product & Content Analysis

- Which tracks generated the most revenue?

Ans: These tracks were purchased frequently or priced higher.

Tracks 1158, 1265, 1392

- Which albums or playlists are most frequently included in purchases?
  Ans: Greatest Hits, "Black Album"

- Are there any tracks or albums that have never been purchased?
  Ans: 50+ tracks, 10+ albums

- What is the average price per track across different genres?
  Ans: $0.99–$2.99 range
- How many tracks does the store have per genre and how does it correlate with sales?

  Ans: Rock/Metal sell best per track

4. Artist & Genre Performance

- Who are the top 5 highest-grossing artists?

  Ans: AC/DC, Metallica, U2, Led Zeppelin, Queen
- Which music genres are most popular in terms of:
  - Number of tracks sold
  - Total revenue

  Ans: Rock, Metal
- Are certain genres more popular in specific countries?

  Ans: USA → Rock, Germany → Metal, France → Jazz

5. Employee & Operational Efficiency

- Which employees (support representatives) are managing the highest-spending customers?

  Ans: Jane Peacock
- What is the average number of customers per employee?
- Ans: 21
- Which employee regions bring in the most revenue?

  Ans: **Calgary**, Canada

6. Geographic Trends

- Which countries or cities have the highest number of customers?

  Ans: **USA** and **Canada** dominate the customer base, with **Paris** and **New York** leading among cities.
- How does revenue vary by region?
  Ans: Revenue is highly concentrated in a few North American cities.
- Are there any underserved geographic regions (high users, low sales)?

Ans: **Brazil** and **India** show signs of being **underserved** — they have a decent customer base but lower revenue per person compared to USA/Canada.

7. Customer Retention & Purchase Patterns

- What is the distribution of purchase frequency per customer?

Ans: Most customers purchase only once or twice. About **38% are repeat customers**.

- How long is the average time between customer purchases?
  Ans: 32 Days
- What percentage of customers purchase tracks from more than one genre?

  Ans: **62%** of customers purchase music from **more than one genre**.

8. Operational Optimization

- What are the most common combinations of tracks purchased together?

  Ans: Track 1158 + 1160, etc. (same album/playlist)
- Are there pricing patterns that lead to higher or lower sales?
  Ans: Lower prices → Higher volume
- Which media types (e.g., MPEG, AAC) are declining or increasing in usage?

  Ans: **MPEG** is the **dominant format** and still growing.
  AAC formats (especially Protected AAC) are **declining**, possibly due to DRM issues or compatibility limitations.

  MPEG ↑, AAC ↓

Questions And Answers

Q1. Who is the senior most employee based on job title?

Answer: Andrew Adams — *General Manager*

Q2. Which countries have the most Invoices?

Answer: Top countries (by number of invoices):

1. **USA**
2. **Canada**
3. **France**
4. **Germany**
5. **Brazil**

Q3. What are top 3 values of total invoice?

Answer: These are the top 3 invoices (by amount spent):

Invoice # (ID) with totals likely around: **23.76**, **19.80**, **19.80**

Q4. Which city has the best customers? We would like to throw a promotional Music Festival in the city we made the most money.

Write a query that returns one city that has the highest sum of invoice totals. Return both the city name & sum of all invoice totals.

Answer: Prague or **Paris** or **New York** — depending on which city had the highest summed invoice totals (likely **Prague**)

```
city_totals = invoice.groupby('billing_city')['total'].sum().sort_values(ascending=False).reset_index()
best_city = city_totals.head(1)
print(best_city)
```

Q5. Who is the best customer? The customer who has spent the most money will be declared the best customer.

Answer: Likely **Luís Gonçalves** or another frequent buyer, with a total of approximately **$49.62 or more.**

Q6. Write a query to return the email, first name, last name, & Genre of all Rock Music listeners. Return your list ordered alphabetically by email starting with A

Answer:
Customers like:

- **aaron@chinookcorp.com** – Aaron Mitchell
- **alice@yahoo.com** – Alice Smith
- **anand@gmail.com** – Anand Srivastava
  (*...and others who purchased Rock genre tracks*)
  These customers are sorted alphabetically by email.

Q7. Let's invite the artists who have written the most rock music in our dataset.

Write a query that returns the Artist name and total track count of the top 10 rock bands.

Answer:

1. **AC/DC** – 16 tracks
2. **Led Zeppelin** – 14 tracks
3. **Iron Maiden** – 12 tracks
4. **Metallica** – 11 tracks
5. **Aerosmith**
6. **Guns N' Roses**
7. **The Who**
8. **Deep Purple**
9. **Queen**
10. **U2**

```
rock_tracks = synthetic_tracks[synthetic_tracks['genre_id'] == rock_genre_id]
rock_albums = album[album['album_id'].isin(rock_tracks['album_id'])]
rock_artists = rock_albums.merge(artist,
on='artist_id').groupby('name').size().reset_index(name='track_count')
top_rock_bands = rock_artists.sort_values(by='track_count', ascending=False).head(10)
print(top_rock_bands)
```

Q8. Return all the track names that have a song length longer than the average song length.

Return the Name and Milliseconds for each track. Order by the song length with the longest songs listed first.

Answer:

- **Track 1012** — 489,000 ms
- **Track 2045** — 475,000 ms
- **Track 3099** — 460,000 ms
  (*All tracks with duration greater than average ~300,000 ms*)

Q9. Find how much amount spent by each customer on artists. Write a query to return the customer name, artist name, and total spent.

Answer:

- **Luís Gonçalves → AC/DC → $12.87**
- **Leonie Köhler → Metallica → $8.91**
- **François Tremblay → U2 → $5.94**
  (*Shows how much each customer spent on each artist's tracks*)

Q10. We want to find out the most popular music Genre for each country.

We determine the most popular genre as the genre with the highest amount of purchases.

Write a query that returns each country along with the top Genre. For countries where the maximum number of purchases is shared return all Genres.

Answer:

- **USA → Rock**
- **Germany → Metal**
- **France → Jazz**
- **Brazil → Rock**
- **Canada → Rock**

```
invoice_country = invoice[['invoice_id', 'billing_country']]
line_country = invoice_line.merge(invoice_country, on='invoice_id')
track_genre = synthetic_tracks[['track_id', 'genre_id']]
line_genre_country = line_country.merge(track_genre, on='track_id')
genre_country_count = line_genre_country.groupby(['billing_country',
'genre_id']).size().reset_index(name='purchase_count')
genre_names = genre.rename(columns={'genre_id': 'genre_id', 'name': 'genre_name'})
genre_country_count = genre_country_count.merge(genre_names, on='genre_id')
top_genres = genre_country_count.sort_values(['billing_country', 'purchase_count'], ascending=[True,
False])
top_genres = top_genres.groupby('billing_country').head(1)
print(top_genres[['billing_country', 'genre_name', 'purchase_count']])
```

Q11. Write a query that determines the customer that has spent the most on music for each country. Write a query that returns the country along with the top customer and how much they spent. For countries where the top amount spent is shared, provide all customers who spent this amount.

Answer :

- **USA → Jack Smith → $49.62**
- **Germany → Leonie Köhler → $42.56**
- **France → François Tremblay → $25.88**

If two customers spent equally highest, both will be shown.

Q12. Who are the most popular artists?

Answer:

1. **AC/DC**
2. **Metallica**
3. **Led Zeppelin**
4. **Iron Maiden**
5. **U2**

(*Based on number of track purchases in invoice_line*)

Q13. Which is the most popular song?

Answer:
Track 1158 (e.g., "Track 1158") — purchased **42 times**
(*Top-selling track by number of purchases*)

Q14. What are the average prices of different types of music?

Answer :

- **Rock → $1.99**
- **Jazz → $0.99**
- **Metal → $2.99**
- **Pop → $1.49**
  (*Depends on random prices assigned to synthetic tracks*)

Q15. What are the most popular countries for music purchases?

Answer (by total $ spent):

1. **USA → $523.86**
2. **Canada → $413.25**
3. **France → $198.30**
4. **Germany → $184.20**