جامعة الملك سعود
King Saud University

كلية علوم الحاسب والمعلومات
قسم تقنية المعلومات

King Saud University
College of Computer and Information Sciences
Department of Information Technology

Spotify®

IT 362 Course Project

Semester-1, 1447H

Student Name	Student ID
Dana Alsalami	443200945
Norah Aljayan	444200832
Aljohara Alsultan	444203635
Layan Alhowaimel	444200969
Alya Almuqren	444200610

Course Instructor: **Dr. Abeer Aldayel**

Entry Logbook

Week	Tasks	Challenges	Tools
Week 2	Topic Selection & Planning Chose our research question: Do artists who span multiple genres have higher popularity compared to those who focus on a single genre? Identified main data sources: Spotify Charts and Spotify Web API.	Due to updates to Spotify's API, we couldn't download the top songs playlists directly so we used Spotify charts as a workaround to get top artists from regional top songs playlists	None (conceptual)
Week 3	Data Collection Downloaded Spotify Charts CSV files from 20 regions (e.g., Global, US, UK, Saudi Arabia, Japan, Brazil). Extracted artist names, track names, chart ranks, streams, weeks on chart, and peak positions. Collected ~1785 unique artists after consolidating the files. Retrieved additional artist information (Spotify ID, genres, popularity score) via Spotify Web API using Spotipy.	Some chart files were downloaded as "404: Not Found" instead of valid CSVs, we to filter out invalid or incomplete files Collecting data for hundreds of artists required a large number of API requests, which often led to rate limits, timeouts, or disconnections. This forced us to implement retries and pauses, which slowed down the data collection process.	requests (for downloading CSVs). os, glob (file handling). spotipy (Spotify API). time (pausing between API calls). spotipy.exceptions, requests.exceptions (error handling).
Week 3	Data Preprocessing Merged multiple CSV files into a single dataset. Removed duplicate artists across charts. Excluded artists without genre information. Stored final dataset as a Pandas DataFrame containing artists with valid genres and popularity scores.	There was incomplete genre information for some artist and duplicate entries across charts, so we had to filter out duplicate entries, as well as excluding artists with empty genre lists, reducing the dataset size but ensuring consistency.	Pandas (merging, cleaning, creating dataframe)
Week 4	Data Analysis & Visualization Created new feature genre_count to classify artists as single-genre or multi-genre. Compared popularity distributions using boxplots. Visualized number of artists in each category with countplots.	Picking between boxplots, histograms, or violin plots for comparing distributions required trial and error to communicate the results clearly.	pandas (feature engineering). matplotlib, seaborn (visualization).
Week 5	Discussion of Results & Biases Identified representation bias (larger markets like US/UK dominating data) and measurement bias (popularity influenced by recent releases). Discussed limitations such as artists with pseudonyms or missing genres.	It was tricky to separate representation bias (market dominance), measurement bias (Spotify's opaque popularity score) Even though we had analyzed around 20 charts with 200 tracks each, the final dataset only has 699 unique artists with a nonempty genre list which raised concerns whether the dataset was large enough	Microsoft word (for report)