Karan Khatke

karan.khatke@gmail.com, karan_khatke@isb.edu | +919179447005, +917987997278 | Indore, India
Linkedin Profile | Github Profile

EDUCATION

LNCT, Indore, India 2016 - 2021
M.Tech. Thermal Engineering (GPA: 8.21/10)
VITS, Indore, India 2010 - 2014
B.E. Mechanical Engineering (GPA: 7.86/10)

SKILLS

Data Engineering & ETL: ETL Pipelines, Data Warehousing, Data Migration, Data Lake Management
Programming & Databases: Python (Pandas, NumPy, Polars), SQL (PostgreSQL, MySQL), Query Optimization, Performance Tuning.
Data Pipeline & Orchestration Tools: Mage AI, Kestra.
Cloud & Infrastructure: Oracle Cloud Infrastructure (OCI), Google Cloud (GCP), Wasabi Cloud Storage
Scraping Skills: Libraries: Requests, BeautifulSoup, Selenium, Playwright.
Data Processing & Automation: API Integration, Data Validation, Data Governance, Real-Time Data Processing.
Data Analytics: Exploratory Data Analysis (EDA), Data Cleaning, Data Visualization (Matplotlib, Seaborn, Plotly), Advanced SQL.
Dashboarding Tools: Tableau, Google Data Studio, Advanced Excel, Libraries: Dash, Streamlit.
Data Science and Machine Learning: Scikit-learn, NLP, LLM, Hugging Face Transformers, Gen AI.
IDE & Team Services: Jupyter Notebooks, Google Collab, VS Code.
Version Control Tools: Git
Other Tools: MS Word and MS PowerPoint.

CERTIFICATIONS

Generative AI with Large Language Models – DeepLearning.AI Oct 2024 – Feb 2025
Courses: Large Language Models (LLMs), Transformer Architecture, Gen AI Life Cycle, Scaling Laws, Model Training. Highlights: 20+ hrs coursework, 3 graded assignments.
Data Engineering Bootcamp – DataExpert.io Nov 2024 – Feb 2025
Courses: Data Modelling, Fact Modelling, Apache Spark, Flink and Kafka, Real-Time Data Processing, Advanced SQL and Analytics, Pipeline Maintenance, KPIs and A/B Testing, Unit Testing PySpark. Highlights: 40+ hrs of coursework, 8 assignments.
Data Analytics Bootcamp, OneLearn.io Apr 2022 - Jun 2022
Courses: Python Programming, Data Analysis & Visualization, SQL & Analytics, Dashboarding and Deployment, Tableau. Highlights: 600+ hrs of coursework, 10 coding assignments, 3 projects (Python, EDA, Pandas, SQL)

WORK EXPERIENCE

Sr. Data Scientist – Bharti Institute of Public Policy, ISB, Mohali (Jan 2024 - Present)

- **Built and optimized ETL pipelines** using **Mage AI**, automating data ingestion from multiple sources for IDP (India Data Portal).
- **Managed cloud-based infrastructure** on **Oracle Cloud (OCI),** ensuring scalability and resource efficiency for high-volume data processing.
- **Developed and maintained data integration workflows** between **CKAN and internal databases,** improving data ingestion speed by 40%.
- **Implemented pipeline performance optimizations,** reducing data processing time by 30% and improving query execution speed.

Research Associate – Bharti Institute of Public Policy, ISB, Mohali (Feb 2023 – Dec 2024)

- **Designed, built, and maintained ETL pipelines** to process **large-scale government datasets**, ensuring efficient data ingestion and transformation.
- **Automated data validation and cleansing workflows**, reducing inconsistencies in **structured and unstructured data** across platforms.
- **Developed static and animated graphics** to visually represent insights derived from the data, effectively communicating key findings and trends to stakeholders.
- Collaborated with data engineers to implement **data quality validation** methods, ensuring structured and accurate datasets on the Himachal Data Portal and Meghalaya Data Portal.
- Stayed updated with emerging trends in **data visualization** and employed innovative approaches to present complex data in a visually appealing and easily understandable manner.

Data Analyst Intern – Amulyam Digital Media Pvt. Ltd., Indore (July 2022 - Dec 2022)

- Analyzed and monitored data-driven campaigns and performance metrics to identify trends and insights. Developed and maintained dashboards and reporting to track KPIs.
- **Integrated REST APIs** to automate data ingestion, reducing manual data collection efforts by 30%.
- Created and optimized **SQL queries** for data retrieval, ensuring fast and reliable access to analytics reports.

Data Analyst Intern – Trendy Dice (Jan 2022 - June 2022)

- **Integrated real-time data updates** for sales tracking using Google Data Studio.
- **Implemented anomaly detection in order deliveries**, preventing 20% of errors in logistics reporting.
- Designed and managed **data validation checks** to detect anomalies in order deliveries, improving reporting accuracy.
- Built an email reminder for the team whenever the delivery time is delayed over 7 days so that escalation can be done for these orders.
- Created weekly reports on Product transactions with respect to products being sold.
- Segmented users using **RFM methodology** for planning customer engagement activity to increase user

Assistant Professor – PIEMR, Indore (Jan 2015 - Feb 2023)

- Analyzing the attendance and results which helps the students to grow and perform well,
- Conveyed subject matter and lecture to the students in a creative way.
- Ensured completion of assigned syllabus within the time frame given.
- Analyzing the previous session data for continuous improvement and NBA accreditation.
- Analyzing and regulating the academic data as per guidelines of governing bodies like AICTE, UGC, accreditation bodies like NBA, NAAC, etc.

Design Consultant – Freelancer (Jan 2014 - Feb 2023)

- Conducting legacy conversions -3D Solid Modeling, Surfacing & Drawing Conversion, and Data Migration - Parametric, Non-Parametric Modeling of Components
- Check the behavior of components on CAE software etc.
- Reviewing the Engineering/design changes & resolving quality related problems associated with the design

PROJECTS

NLP Text Analytics & Sentiment Engine (link)

Tools used: Python, NLTK, Scikit-Learn, BeautifulSoup, WordCloud, Pandas, Matplotlib, Seaborn

- Built a full **NLP pipeline** to scrape, clean, and analyse 113+ articles, generating sentiment, readability, and linguistic metrics.

- Implemented **lexicon-based sentiment scoring, readability models** (Fog Index, complex words), and **TF-IDF-based topic extraction**.
- Applied **clustering (K-Means)** for article grouping and created comprehensive visual reports including sentiment distribution, correlation matrix, and word clouds.
- Delivered a production-style pipeline generating **structured Excel outputs and automated HTML insights reports**.

LGD Mapping Application (link)

Tools used: Python, RapidFuzz, Pandas, CLI Automation, Logging, Data Validation

- Built a **hierarchical entity-matching engine** using name normalization, exact UID matching, and **fuzzy ML similarity scoring (95%/90%)** to reconcile district–block–village data with LGD (Local Government Directory) codes.
- Automated **multi-level mapping (3–5 levels), parent-aware matching**, and chunked large-scale processing with comprehensive logs and quality reports.
- Added full **data quality validation**, error handling, duplicate detection, and chunk processing for large datasets, ensuring reliability for government-scale data volumes.
- Implemented comprehensive **summary reports**, detailed logs, and quality metrics (match percentages, confidence levels, alternative suggestions) for auditability and debugging.

Fuel & Fleet Trends – India's Transition to EV (link)

Tools used: Google Cloud Platform (GCP) and Storage (GCS), Kestra, Terraform, BigQuery, dbt, and Looker Studio

- Developed an end-to-end data pipeline to monitor and visualize the adoption of electric vehicles (EVs) in India.
- Implemented web scraping to collect vehicle registration data across various fuel types and categories.
- Utilized dbt for data transformation and modeling, and Kestra for orchestrating workflows.
- Deployed infrastructure using Terraform on GCP, ensuring scalability and reliability.
- Created dashboards to provide insights into EV trends, aiding stakeholders in understanding the transition dynamics.

Exploratory Data Analysis and Visualization, Deployment of an interactive dashboard of Movie Analytics.

Tools used: Pandas, Python, Matplotlib, Seabourn, Plotly, and Dash libraries

- These analyses are helpful for movie industries by considering movies preferred by the audience and movies have high earning potential.
- Worked with a dataset of 85,855 movies of the Imdb rating system and 17712 movies of the Tomato meter rating system collected from Kaggle.
- Carried out data interpretation, cleaning, and modification to prepare the data for further analysis.
- Through EDA and visualization, one of the vital insights is drawn that the Animation genre movies have the highest earning potential and higher ratings in both rating systems.

Exploratory Data Analysis and Visualization of New York Yellow Taxi Dataset. (link)

Tools used: Pandas, Python, Matplotlib, and Seaborn libraries

- The analysis is done to see the impact of the pandemic on the yellow taxi business in New York.
- Worked with the trip record data for February 2020 (pre-pandemic) of 6.2M data and for June 2020 (post-pandemic) of 0.5M data obtained from the TLC trip record data New York Taxi dataset.
- Carried out data interpretation, cleaning, and modification to prepare data for work.
- One of the clear insights from the data is that the number of trips in the post-pandemic period (June 2022) is significantly reduced by around 93% as compared to the pre-pandemic period (February 2022).

Email Automation uses Python program for reminder mail of scheduled classes. (link)

Tools used: VS Code, Smtplib, Windows Task Scheduler

- Create an email reminder program for the students of a course series that sends a reminder email to the students just before the class day. Create a Python script that reads a text file containing a class schedule.
- Send the mail to senders using a script file by using the Smtplib library(using VS-code)
- Script file automation is carried out through Windows Task Scheduler.

Video Game Data Analytics SQL project (link)

Tools used: PostgreSQL, Python, Pandas, Jupyter Notebook, Pyscopg2-binary, and the sqlalchemy library

- A real-time videogame dataset obtained from vgchartz.com is used in the project to serve the video game industry.
- The ETL (Extract, Transform, and Load) technique is used in this project. After that, analysis is conducted using the PostgreSQL SQL programming tool.
- Developing an SQL script to create tables in PostgreSQL in accordance with the ERD diagram 's-built structure. Dividing the altered data in accordance with the above-mentioned SQL schema.
- Pyscopg2-binary and the sqlalchemy library package are used to connect to PostgreSQL from a Jupyter notebook and dump data from data frames into the appropriate columns.