R Data Visualization Project Booklet (20 Detailed Projects)

Project 1: COVID-19 Trend Analysis in India

Problem Statement: Analyze the progression of COVID-19 cases in India to identify surges, recovery trends, and testing impact over time.

Description: Using official data sources, create visualizations to display the evolution of confirmed, recovered, and death cases across Indian states over time. Compare lockdown periods and testing rates with case trends.

Dataset: - Source: https://www.kaggle.com/datasets/imdevskp/covid-19-in-india

Constraints:

- Use time series plots with daily and cumulative trends.

- Include stacked bar charts for state-wise comparisons.

- Must include one interactive plot using `plotly`.

Viva Questions:

1. What does the cumulative line graph tell us about the spread?

2. Why did you use stacked bars for state-wise analysis?

3. How did you manage missing or inconsistent data?

Project 2: Air Quality Monitoring Across Cities

Problem Statement: Visualize air quality metrics across Indian cities and assess seasonal or geographic variations.

Description: Use AQI values to compare pollutant levels like PM2.5 and NO2 across different cities. Detect anomalies during lockdowns or festival seasons.

Dataset: - Source: https://www.kaggle.com/datasets/sagnik1511/air-quality-data-in-india

Constraints:

- Include box plots for seasonal comparisons.

- Must visualize PM2.5 and PM10 separately.

- Include a leaflet map showing average AQI per city.

Viva Questions:

1. What does the box plot reveal about winter air quality?

2. Why is mapping AQI on a leaflet map effective?

3. How did you handle missing pollutant readings?

Project 3: Retail Sales Dashboard (BigMart)

Problem Statement: Develop a dashboard to visualize item sales performance by product category and outlet characteristics.

Description: Use BigMart sales data to understand how item categories perform across locations, outlet types, and pricing levels.

Dataset: - Source: https://www.kaggle.com/datasets/brijbhushannanda1979/bigmart-sales-data

Constraints:

- Use bar plots for product-wise sales.

- Create box plots for sales vs MRP.

- Interactive filter by outlet type using `plotly`.

Viva Questions:

1. What sales trend did you find most surprising?

2. Which visualization showed product performance most clearly?

3. How did you handle skewed data in MRP?

Project 4: Population Growth Visualization by Country

Problem Statement: Compare population trends across top 10 countries over the last 50 years.

Description: Visualize long-term changes in population using both growth lines and choropleth maps to highlight top-growing nations.

Dataset: - Source: https://data.worldbank.org/indicator/SP.POP.TOTL

Constraints:

- Use line plots for time series.

- Create a choropleth using `leaflet` or `tmap`.

- Normalize population per million for comparison.

Viva Questions:

1. Why normalize population for comparison?

2. What growth patterns are visible over time?

3. How did you join tabular data with geographic shapes?

Project 5: Student Performance Visualization

Problem Statement: Analyze student marks distribution by subject, gender, and school to uncover performance trends.

Description: Create a dashboard comparing average marks across different subjects, schools, and demographic factors.

Dataset: - Source:
https://www.kaggle.com/datasets/spscientist/students-performance-in-exams

Constraints:

- Use box plots for subject-wise performance.

- Use grouped bar plots for school-wise and gender-wise comparisons.

- Must include a histogram for overall performance.

Viva Questions:

1. How did you analyze subject difficulty from the plots?

2. Which visualization revealed gender-based trends clearly?

3. What insights can the school admin draw from your visuals?

Project 6: Crime Rates by State

Problem Statement: Visualize various categories of crime reported in Indian states to understand regional patterns.

Description: Present comparisons across states for crimes like theft, assault, and cybercrime using bar charts and maps.

Dataset: - Source: https://data.gov.in/

Constraints:

- Use a choropleth map for total crime rate.

- Use stacked bar plots for crime categories.

- Compare urban vs rural using faceted plots.

Viva Questions:

1. Why is mapping crime rates more insightful than just plotting values?

2. What category of crime showed the largest regional difference?

3. How did you highlight outliers?

Project 7: IPL Player Performance Analysis

Problem Statement: Analyze and visualize Indian Premier League (IPL) player statistics to evaluate consistency and value.

Description: Visualize batting and bowling stats such as runs, wickets, economy rate, and strike rate.

Dataset: - Source: https://www.kaggle.com/datasets/patrickb1912/ipl-complete-dataset-2008-2020

Constraints:

- Use scatter plots for strike rate vs average.

- Use bar plots for team-wise contributions.

- Must use radar chart for all-rounders.

Viva Questions:

1. How did you identify consistent performers?

2. Why is a radar chart useful for all-rounders?

3. What visual cues indicate match impact?

Project 8: Traffic Accident Analysis

Problem Statement: Visualize and analyze traffic accident patterns based on time, location, and severity.

Description: Use historical accident data to identify hotspots, time trends, and casualty patterns. Focus on seasonal, weekly, and hourly distributions.

Dataset: - Source: https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

Constraints:

- Use heatmaps for hourly and weekly frequency.

- Display top 10 accident-prone cities using bar plots.

- Show accident severity using pie or donut charts.

Viva Questions:

1. What time of day sees most accidents?

2. How do weather and visibility relate to severity?

3. Why are heatmaps effective for time analysis?

Project 9: Financial Market Visualization (Stock Prices)

Problem Statement: Analyze stock price trends of major companies and compare returns over a year.

Description: Use Yahoo Finance or NSE/BSE stock data to visualize trends in closing prices, moving averages, and relative volume.

Dataset: - Source: Use `quantmod` or download from Yahoo Finance via `tidyquant`

Constraints:

- Use line plots for stock price trends.

- Use candlestick charts to show open-high-low-close (OHLC).

- Plot moving averages with trend lines.

Viva Questions:

1. What does a candlestick represent?

2. How do moving averages help in trend prediction?

3. What patterns signal bullish/bearish trends?

Project 10: Tourism Inflow by Country

Problem Statement: Visualize tourism inflow trends over the past decade to top global destinations.

Description: Identify countries with rising tourism numbers and seasonal patterns using year and month-wise visitor data.

Dataset: - Source: https://data.worldbank.org/indicator/ST.INT.ARVL

Constraints:

- Use area charts for cumulative growth.

- Use bar plots for yearly ranking.

- Visualize seasonality using line plots per country.

Viva Questions:

1. Which countries showed the fastest growth in tourism?

2. Why use area chart over bar chart in this case?

3. What months show peaks and troughs?

Project 11: Global Hunger Index Visualization

Problem Statement: Analyze the Global Hunger Index and compare countries based on nutrition and undernourishment indicators.

Description: Compare GHI scores, child mortality, underweight children stats using multiple charts across years.

Dataset: - Source: https://www.globalhungerindex.org/download/all.html

Constraints:

- Use world map for GHI scores.

- Use line plots to show progress over time.

- Use lollipop plots for country-wise comparison.

Viva Questions:

1. What indicators contribute to the GHI?

2. How has India progressed over the years?

3. Why choose a map for comparing countries?

Project 12: Movie Ratings & Genre Popularity (IMDB)

Problem Statement: Analyze movie ratings and identify the most successful genres by decade.

Description: Use IMDB or TMDB datasets to compare genre-wise average ratings and box office revenue trends.

Dataset: - Source: https://www.kaggle.com/datasets/PromptCloudHQ/imdb-data

Constraints:

- Use bar plots for genre popularity.

- Use scatter plots for rating vs box office.

- Use heatmaps for ratings by genre and decade.

Viva Questions:

1. Which genre performs best over time?

2. How do ratings vary with budget or revenue?

3. Why did you choose a heatmap for decade analysis?

Project 13: Ecommerce Customer Segmentation

Problem Statement: Visualize and segment online shoppers based on spending behavior and purchase frequency.

Description: Use RFM analysis (Recency, Frequency, Monetary) to visualize customer clusters.

Dataset: - Source: https://www.kaggle.com/datasets/carrie1/ecommerce-data

Constraints:

- Use k-means clustering to create customer segments.

- Visualize clusters using 2D/3D scatter plots.

- Use bar plots for segment-wise value.

Viva Questions:

1. What is RFM analysis and how is it visualized?

2. How do visual clusters assist marketing?

3. How did you decide number of segments?

Project 14: Road Transport Visualization in India

Problem Statement: Analyze the trends in vehicle registrations and fuel type over the last decade in India.

Description: Use transport ministry data to understand vehicle growth, fuel type adoption, and electric vehicle uptake.

Dataset: - Source: https://data.gov.in/

Constraints:

- Use stacked area plots for fuel type.

- Use bar plots for state-wise vehicle registration.

- Must visualize EV growth using line chart.

Viva Questions:

1. Which states showed highest EV growth?

2. Why is a stacked area chart suited for this?

3. How did you adjust for missing state data?

Project 15: Hospital Beds and Health Infrastructure

Problem Statement: Visualize the distribution and growth of healthcare facilities across Indian states.

Description: Compare hospital beds, primary health centers, and rural vs urban distribution.

Dataset: - Source: https://data.gov.in/catalog/stateut-wise-health-infrastructure

Constraints:

- Use bar charts and maps.

- Normalize beds per 1000 people.

- Use faceted plots for rural vs urban.

Viva Questions:

1. What visual showed regional disparity best?

2. Why normalize bed count?

3. What does faceting add to your analysis?

Project 16: Global Temperature Change

Problem Statement: Visualize global warming trends across continents and key cities.

Description: Use climate datasets to showcase temperature rise, anomalies, and decadal averages.

Dataset: - Source: https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data

Constraints:

- Use line charts and heatmaps.

- Use maps to show temperature anomalies.

- Use smooth trend lines.

Viva Questions:

1. What data transformations did you apply?

2. Why is anomaly visualization important?

3. How did you address global vs local trends?

Project 17: Educational Expenditure Visualization

Problem Statement: Visualize government and private expenditure on education in India across years.

Description: Compare expenditure by level (primary, secondary, higher) and state.

Dataset: - Source: https://data.gov.in/

Constraints:

- Use bar charts and trend lines.

- Normalize data per capita or per student.

- Use faceted charts for level-wise analysis.

Viva Questions:

1. Which state invests most per student?

2. How did you compare across levels?

3. What visualization best shows trends?

Project 18: Water Resource Usage

Problem Statement: Analyze and visualize how water resources are used by sector and region.

Description: Use ministry of water resources data to identify water demand trends.

Dataset: - Source: https://data.gov.in/

Constraints:

- Use pie charts for sectoral use.

- Use bar plots for state-wise availability vs demand.

- Show shortages using diverging bar charts.

Viva Questions:

1. Why use diverging bars for shortages?

2. What sector consumes most water?

3. How does visualizing supply vs demand help?

Project 19: Startup Funding Trends

Problem Statement: Visualize funding activity in Indian startups by city, sector, and year.

Description: Use startup funding data to assess trends, unicorns, and investor activity.

Dataset: - Source: https://www.kaggle.com/datasets/ashishjangra/indian-startup-funding

Constraints:

- Use bar charts, treemaps.

- Timeline chart for funding trends.

- Highlight unicorns with annotations.

Viva Questions:

1. What sectors got most funding?

2. Why did you choose a treemap?

3. What is the advantage of using timelines?

Project 20: Electricity Consumption Across States

Problem Statement: Visualize electricity usage across states and sectors.

Description: Compare state-wise consumption patterns, rural vs urban usage, and industrial loads.

Dataset: - Source: https://data.gov.in/

Constraints:

- Use stacked bar plots.

- Use maps for per capita usage.

- Include trend lines for years.

Viva Questions:

1. Which state has highest per capita use?

2. How did you split rural vs urban?

3. How did your visuals highlight peak demand areas?

Project 21: Electric Vehicle Adoption Trends

Problem Statement: Visualize the growth in electric vehicle registrations across states and over time.

Description: Analyze EV registration trends using government or open datasets. Compare adoption between urban vs rural and vehicle types.

Dataset:

- Source: https://data.gov.in/

Constraints:

- Line plots for time trends.
- Pie charts for vehicle type distribution.

- Leaflet map for state-wise density.

Viva Questions:

1. How did urban areas differ from rural in adoption?
2. Why use pie charts despite their limitations?
3. How did you normalize population across states?

Project 22: Telecom User Consumption Visualization

Problem Statement: Visualize telecom usage behavior in terms of call, SMS, and data usage over user segments.

Description: Use telecom data to analyze usage by prepaid/postpaid users and across weekdays vs weekends.

Dataset:

- Source: https://www.kaggle.com/datasets/muthuj7/mobile-telecom-customer

Constraints:

- Faceted bar charts for usage type.
- Box plots for outliers in data consumption.
- Heatmaps for time-of-day usage patterns.

Viva Questions:

1. What time slot shows the highest traffic?
2. What is the advantage of facet wrapping in ggplot2?
3. How did you deal with missing plan details?

Project 23: Online Education Participation Trends

Problem Statement: Analyze and visualize trends in online course enrollments and completions.

Description: Use data from MOOCs or learning platforms to explore learner demographics and success rates.

Dataset:

- Source: https://www.kaggle.com/datasets/jessemostipak/udemy-courses-dataset

Constraints:

- Bar plots for course enrollments by subject.
- Line plots for time-based trends.
- Word cloud for most common course titles.

Viva Questions:

1. What factors impacted course completion?
2. How does subject popularity vary by price?
3. Why use a word cloud in education data?

Project 24: Fitness Tracker Data Visualization

Problem Statement: Explore activity levels, steps, and calorie burn patterns using fitness tracker data.

Description: Visualize activity trends across time and compare weekdays vs weekends.

Dataset:

- Source: https://www.kaggle.com/datasets/arashnic/fitbit

Constraints:

- Line chart for daily steps.
- Heatmap for hourly activity.
- Radar chart for weekly summaries.

Viva Questions:

1. What trends do you observe in sleep vs steps?
2. How do radar charts help compare multiple activities?
3. What visualization helped spot inactive periods?

Project 25: YouTube Trending Video Analysis

Problem Statement: Analyze viewer behavior and content categories on YouTube trending videos.

Description: Visualize likes, views, and comment patterns across content categories.

Dataset:

- Source: https://www.kaggle.com/datasets/datasnaek/youtube-new

Constraints:

- Bar plot of top 10 most viewed videos.
- Violin plots for likes distribution.
- Word cloud of trending video titles.

Viva Questions:

1. What category trended most frequently?
2. How did likes correlate with comments?
3. Why use violin plots instead of box plots?

Project 26: Retail Store Customer Traffic Analysis

Problem Statement: Visualize customer visit patterns to retail stores based on time and demographics.

Description: Use footfall and transaction data from retail chains to understand high-traffic hours, demographics, and store comparisons.

Dataset:

- Source: https://www.kaggle.com/datasets/kyanyoga/sample-sales-data

Constraints:

- Line charts for hourly/daily footfall.
- Bar charts for store-wise customer count.
- Density plot for age distribution.

Viva Questions:

1. What trends did you observe during weekends?
2. How does customer traffic vary with time of day?
3. Why use density plots for age visualization?

Project 27: Spotify Song Popularity Visualization

Problem Statement: Analyze how audio features of songs correlate with popularity.

Description: Visualize the popularity of songs based on tempo, danceability, loudness, and other musical features.

Dataset:

- Source: https://www.kaggle.com/datasets/geomack/spotifyclassification

Constraints:

- Correlation heatmap of all audio features.
- Scatter plots between popularity vs features.
- Histogram of popularity scores.

Viva Questions:

1. Which features correlate most with popularity?
2. What patterns do you observe in danceability and tempo?
3. Why is heatmap used in feature analysis?

Project 28: Mental Health and Work Analysis

Problem Statement: Visualize patterns in mental health issues related to workplace settings.

Description: Explore how company size, remote work policy, and gender influence mental health outcomes.

Dataset:

- Source: https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey

Constraints:

- Grouped bar plots for company policy impact.
- Box plots for age and mental health treatment.
- Donut chart for gender distribution.

Viva Questions:

1. What workplace factor most affects mental health?

2. How did you manage categorical variables?
3. Why choose a donut over a pie chart?

Project 29: Climate Change Indicators Visualization

Problem Statement: Explore how climate indicators (temperature, CO2 levels, sea level) have changed over decades.

Description: Create a dashboard-style analysis of climate trends globally using environmental datasets.

Dataset:

- Source: https://datahub.io/core/global-temp

Constraints:

- Line plots for temperature and CO2 over time.
- Dual-axis chart for temp vs CO2.
- Map showing country-wise temperature anomalies.

Viva Questions:

1. What regions show the most drastic changes?
2. Why use a dual-axis chart?
3. How did you convert year columns into time-series?

Project 30: Employee Attrition and Performance Analysis

Problem Statement: Visualize the impact of performance metrics and job satisfaction on attrition.

Description: Use HR analytics data to discover patterns related to employee attrition and satisfaction.

Dataset:

- Source: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

Constraints:

- Bar charts for attrition by department.
- Violin plots for job satisfaction vs attrition.
- Heatmap for performance score vs attrition.

Viva Questions:

1. What department had the highest attrition?
2. How are violin plots different from box plots?
3. What performance metric had the most influence?

Project 31: Loan Approval Visualization

Problem Statement: Visualize trends and patterns influencing loan approval decisions.

Description: Explore how factors like income, loan amount, credit history, and marital status affect loan approval rates.

Dataset:

- Source: https://www.kaggle.com/datasets/altruistdelhite04/loan-prediction-problem-dataset

Constraints:

- Bar plots for approval vs applicant characteristics.
- Box plots for income vs loan amount.
- Pie chart for overall approval rate.

Viva Questions:

1. What feature showed the biggest difference in approval rates?
2. Why did you choose box plots for income comparisons?

3. How did you handle missing data in this dataset?

Project 32: Online Education Performance Visualization

Problem Statement: Analyze student performance trends in online courses.

Description: Use engagement and performance metrics to evaluate how students are performing in virtual learning environments.

Dataset:

- Source: https://www.kaggle.com/datasets/aljarah/xAPI-Edu-Data

Constraints:

- Histogram for grade distributions.
- Heatmap for correlation between study time and grade.
- Grouped bar chart for gender vs performance level.

Viva Questions:

1. Which gender had a higher proportion of high performers?
2. What does the heatmap reveal about study time?
3. What other visualization could represent this data well?

Project 33: Natural Disaster Frequency by Country

Problem Statement: Visualize frequency and type of natural disasters across the globe.

Description: Analyze global disaster data to observe which regions are most affected and by what types of disasters.

Dataset:

- Source: https://www.kaggle.com/datasets/emdat/emdat-data-on-natural-disasters

Constraints:

- Choropleth map for frequency by country.
- Bar chart for disaster types.
- Timeline of major events.

Viva Questions:

1. Which region had the highest number of disasters?
2. Why did you choose a choropleth map for this?
3. What does the timeline help us interpret?

Project 34: Air Travel Patterns and Delay Visualization

Problem Statement: Visualize trends in flight routes, delays, and airline performance.

Description: Explore air travel data for delay trends, popular routes, and seasonal travel peaks.

Dataset:

- Source: https://www.kaggle.com/datasets/usdot/flight-delays

Constraints:

- Line plots for monthly delays.
- Map for popular flight paths.
- Box plot for airline-wise delay distribution.

Viva Questions:

1. What month showed the most delays?
2. How did you represent spatial flight routes?
3. Which airline had the least consistent performance?

Project 35: Hospital Admissions and Resource Utilization

Problem Statement: Visualize how hospitals allocate resources like beds, ICUs, and staff across departments.

Description: Analyze hospital admission data and visualize bed occupancy and departmental pressure.

Dataset:

- Source: https://www.kaggle.com/datasets/joniarroba/noshowappointments

Constraints:

- Stacked bar charts for department-wise admissions.
- Heatmap for day-wise occupancy.
- Donut chart for department usage.

Viva Questions:

1. What day of the week had the highest load?
2. How did visualizations help compare departments?
3. What improvements can you suggest for hospital management?

Project 36: Online Product Review Analysis

Problem Statement: Visualize product ratings and review trends from e-commerce platforms.

Description: Use user review data to analyze product sentiment, ratings distribution, and word frequency.

Dataset:

- Source: https://www.kaggle.com/datasets/datafiniti/consumer-reviews-of-amazon-products

Constraints:

- Bar chart for average rating by category.
- Word cloud of most used terms.
- Line chart of review volume over time.

Viva Questions:

1. What product category had the most reviews?
2. How did the word cloud help in sentiment analysis?
3. What challenges did you face with text data?