MULTI-DOMAIN DATA ANALYTICS PORTFOLIO

==========

ABSTRACT

This report presents a multi-domain data analytics portfolio consisting of five distinct data analysis projects across different application areas: Retail, Education, Weather, Healthcare, and Social Media. The objective of this portfolio is to demonstrate practical data analytics skills including data cleaning, transformation, visualization, and insight generation using Python. Pandas was used for data manipulation, while Matplotlib was used for creating visualizations. Each project extracts meaningful insights from real-world or simulated datasets and presents them in a structured and interpretable manner.

TOOLS AND TECHNOLOGIES USED

- Python 3.x: Core programming language for analysis
- Pandas: Data cleaning, manipulation, and aggregation
- NumPy: Numerical operations
- Matplotlib: Data visualization
- Jupyter Notebook: Interactive development and analysis
- VS Code: Code editing and project organization

PROJECT 1: SUPERMARKET SALES ANALYSIS (RETAIL)

Objective:

The goal of this project is to analyze supermarket sales data to identify sales trends, top-performing products, and branch-wise performance.

Dataset Description:

The dataset contains transaction-level sales data including date, product line, branch, payment method, quantity, and total sales amount.

Methodology:

- Cleaned and validated data using Pandas

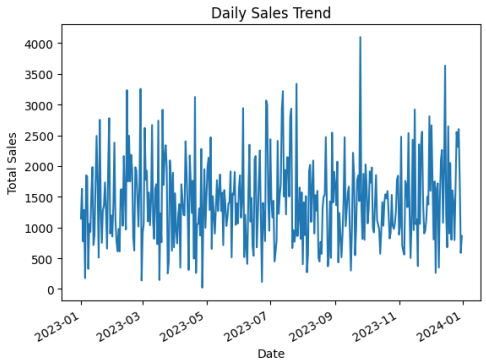
- Aggregated sales using groupby operations

- Analyzed trends over time

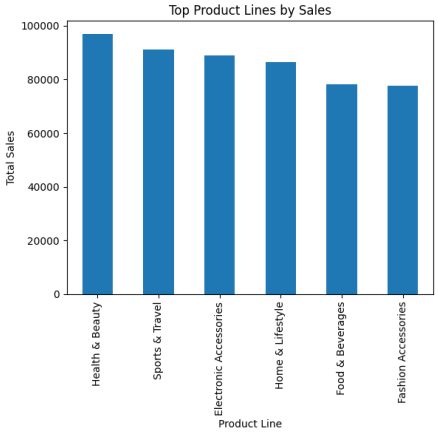
- Created bar charts and line plots for visualization

Visualizations:

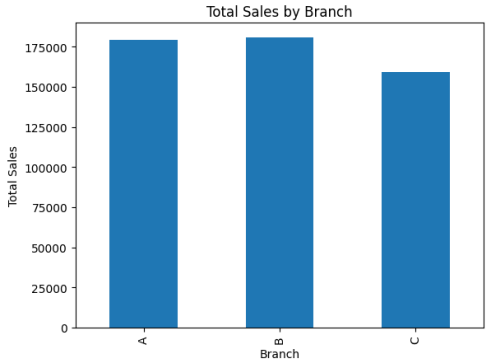
- Daily Sales Trend



- Top Product Lines by Sales



- Branch-wise Sales



Business Insights:

- Certain product lines contribute significantly to total revenue

- Sales vary across branches indicating performance differences

- Sales show noticeable variation across days

PROJECT 2: STUDENT PERFORMANCE ANALYSIS (EDUCATION)

Objective:

To analyze student academic performance and identify factors influencing success and failure.

Dataset Description:

The dataset includes student demographic details, subject scores, and test preparation status.

Methodology:

- Calculated average scores across subjects

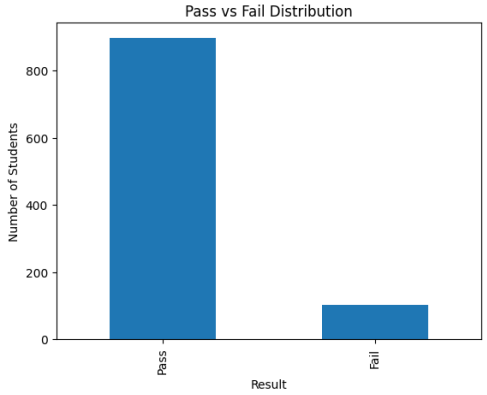
- Classified students into pass/fail categories

- Grouped data based on gender and preparation course

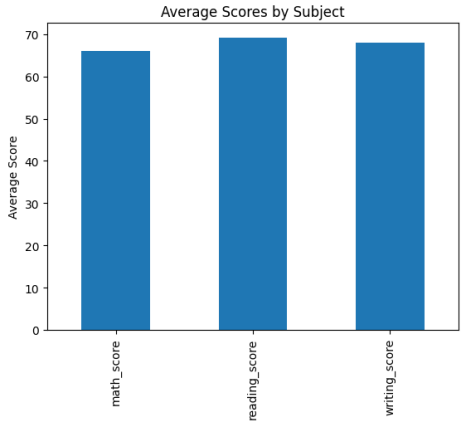
- Visualized academic trends

Visualizations:

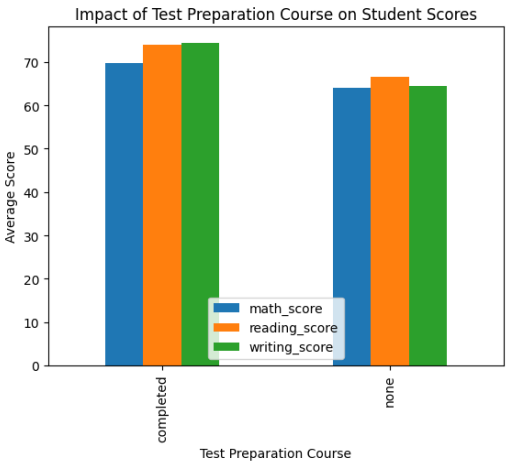
- Pass vs Fail Distribution



- Average Scores by Subject



- Impact of Test Preparation Course



Insights:

- Students who completed test preparation courses performed better

- Reading scores were higher compared to math scores

- Gender-based performance differences were observed

PROJECT 3: WEATHER DATA ANALYSIS

Objective:

To analyze weather patterns and identify relationships between temperature, rainfall, and humidity.

Dataset Description:

The dataset contains meteorological measurements such as temperature, rainfall, humidity, and pressure.

Methodology:

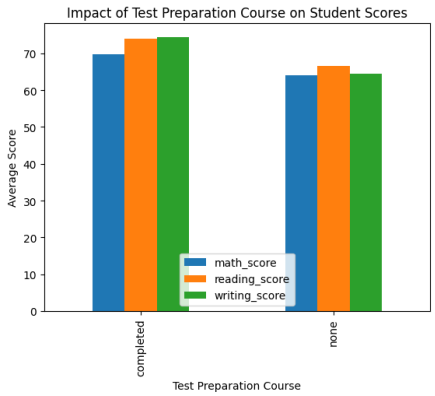
- Statistical analysis using mean and quantiles

- Identification of extreme temperature events

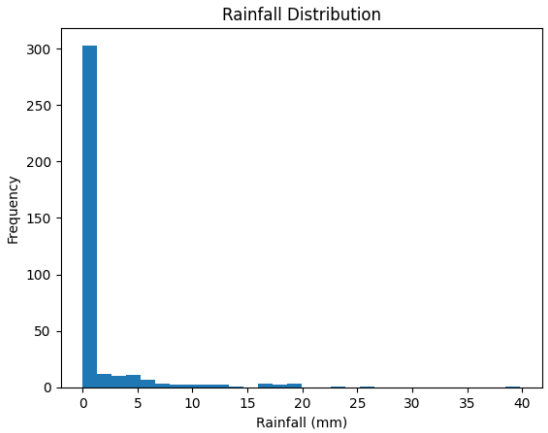
- Relationship analysis between weather variables

Visualizations:

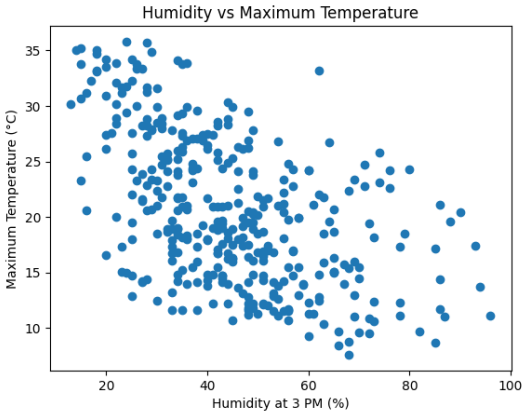
- Maximum Temperature Distribution



- Rainfall Distribution



- Humidity vs Maximum Temperature



Insights:

- Rainfall distribution is highly skewed

- Higher humidity generally corresponds to higher temperatures

- Extreme temperature events occur infrequently but are significant

PROJECT 4: COVID-19 DATA ANALYSIS (HEALTHCARE)

Objective:

To analyze global COVID-19 data and compare country-level impact and mortality rates.

Dataset Description:

The dataset includes country-wise COVID-19 statistics such as confirmed cases, deaths, recoveries, and active cases.

Methodology:

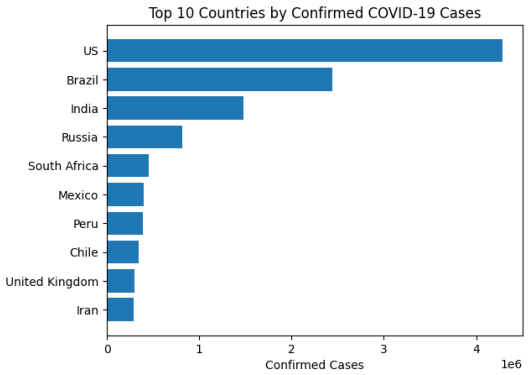
- Sorting and filtering top affected countries

- Calculating mortality-related metrics

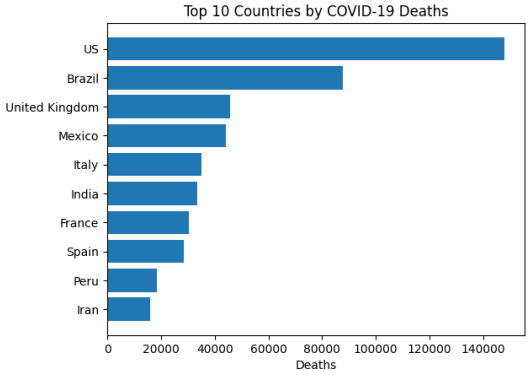
- Comparative analysis across countries

Visualizations:

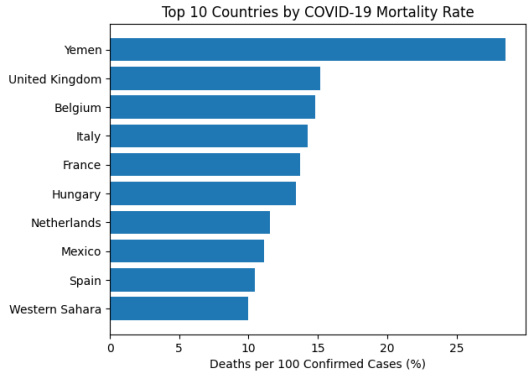
- Top 10 Countries by Confirmed Cases



- Top 10 Countries by Deaths



- Mortality Rate Analysis



Insights:

- Some countries show disproportionately high mortality rates

- Active case counts indicate healthcare system burden

- COVID impact varies significantly across regions

PROJECT 5: INSTAGRAM ENGAGEMENT ANALYSIS (SOCIAL MEDIA)

Objective:

To analyze Instagram engagement metrics and understand content performance and audience behavior.

Dataset Description:

The dataset includes impressions, likes, comments, shares, saves, and traffic source information.

Methodology:

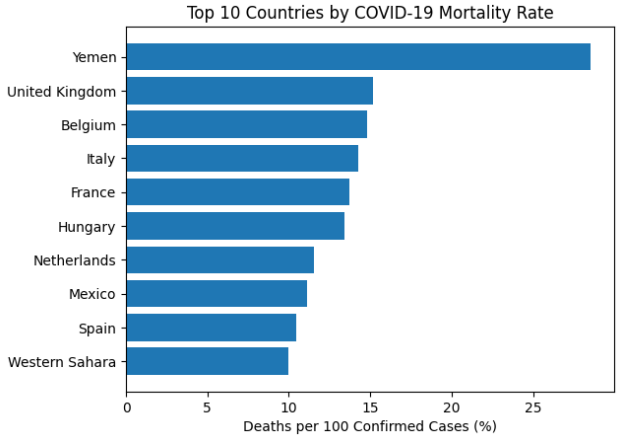
- Computed engagement metrics

- Correlation analysis between engagement variables

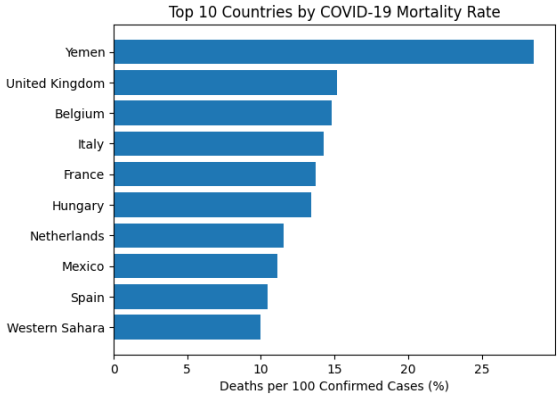
- Traffic source contribution analysis

Visualizations:

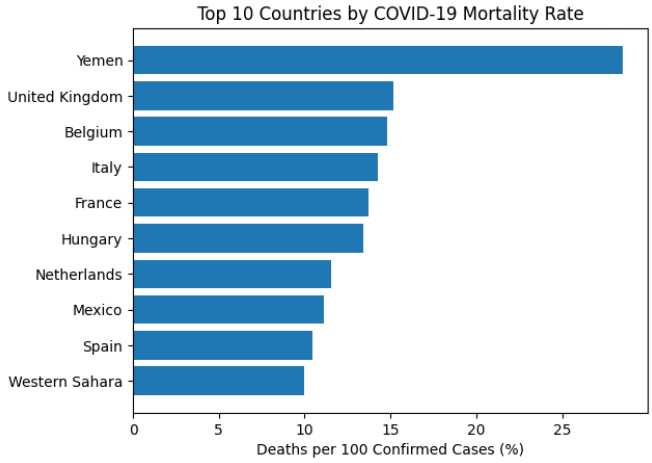
- Engagement Rate Distribution



- Likes vs Comments



- Traffic Source Contribution



Insights:

- Posts with higher likes generally receive more comments

- Hashtags significantly increase content reach

- Saves indicate strong user interest in content

TECHNICAL IMPLEMENTATION DETAILS

- Pandas was used for data loading, cleaning, aggregation, and transformation

- GroupBy operations were used extensively for analytical summaries

- Matplotlib was used to generate bar charts, histograms, scatter plots, and line graphs

- Project structure was organized into data, notebooks, visualizations, and documentation folders

TESTING AND VALIDATION

Data validation was performed to ensure accuracy and consistency across all projects. Column names, data types, and non-null values were verified using Pandas inspection methods. The df.info() output confirms correct data loading, appropriate data types, and absence of unexpected missing values. Aggregation results were cross-checked using descriptive statistics to validate analytical calculations. Screenshots of validation outputs have been included as evidence.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 14 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   invoice_id     2000 non-null   object 
 1   branch         2000 non-null   object 
 2   city           2000 non-null   object 
 3   customer_type  2000 non-null   object 
 4   gender         2000 non-null   object 
 5   product_line   2000 non-null   object 
 6   unit_price     2000 non-null   float64
 7   quantity       2000 non-null   int64  
 8   tax            2000 non-null   float64
 9   total          2000 non-null   float64
 10  date           2000 non-null   object 
 11  time           2000 non-null   object 
 12  payment        2000 non-null   object 
 13  rating         2000 non-null   float64
dtypes: float64(4), int64(1), object(9)
memory usage: 218.9+ KB
```

CHALLENGES AND SOLUTIONS

- Encoding errors in CSV files were resolved by specifying appropriate encodings
- Column name mismatches were fixed through inspection and correction
- File path issues were resolved using relative paths

CONCLUSION

This multi-domain data analytics portfolio demonstrates the ability to analyze diverse datasets, apply analytical techniques, and extract meaningful insights. The projects highlight strong foundations in data analysis, visualization, and problem-solving using Python-based tools.

FUTURE ENHANCEMENTS

- Integration of machine learning models
- Interactive dashboards using Power BI or Tableau
- Automation of data pipelines

REFERENCES

- Pandas Documentation
- Matplotlib Documentation
- Kaggle Datasets