REPORT

Project C12 - Crime rates in LA, Chicago, and Portland compared to Vancouver

Team members - Andreas Kelder, Richard Mario Raun

Task 1. Setting up

Link to repository: https://github.com/RichardMarioRaun/IDS-crime-data

Task 2. Business understanding

Background

Crime rates are a key indicator of safety in urban areas. Comparing Vancouver, Canada, with U.S. cities like Portland, Los Angeles (LA), and Chicago offers valuable insights into how urban policies, law enforcement, and socioeconomic factors influence crime. This analysis focuses on crime rates, severity, and seasonal or time-of-day fluctuations using datasets from these cities.

Business Goals

1. **Determine Vancouver's Overall Crime Rate in Comparison to U.S. Cities:**
   Analyze whether Vancouver experiences lower overall crime rates than Portland, LA, and Chicago.
2. **Assess Crime Severity Across Cities:**
   Understand the relative severity of crimes in each city to identify significant patterns or areas of concern.
3. **Examine Seasonal or Time-of-Day Crime Patterns:**
   Investigate how crime rates vary by season or time of day to inform community safety strategies.

Business Success Criteria

The study is deemed successful if it provides actionable insights into:

- The relative safety of Vancouver compared to U.S. cities.
- Key differences in crime severity and temporal patterns.
- Recommendations for resource allocation based on findings.

Assessing the Situation

Inventory of Resources

- **Datasets:**
    - The datasets include crime data from Vancouver, Los Angeles, Chicago, and Portland, sourced from Kaggle. The largest dataset (Los Angeles and Chicago) is approximately 1.98 GB, with Portland and Vancouver datasets being 84.64 MB and 82.12 MB respectively.
- **Tools & Expertise:**
    - Analytical tools: Python and it's libraries like (numpy, seaborn,scikit-learn, matplotlib) for data cleaning, transformation, and visualization.

Requirements, Assumptions, and Constraints

- Datasets are assumed to have consistent crime categories or can be mapped to a standard schema.
- Crime severity is quantified using weighted indices, e.g., property damage vs. violent crime.
- Timestamps, if present, are analyzed in local time.
- Constraints include potential data gaps or differences in reporting standards between countries.

Risks and Contingencies

- **Risk:** Variability in data quality across cities.
  Contingency: Cross-validation of categories to ensure comparability.
- **Risk:** Inconsistent time periods covered in datasets.
  Contingency: Normalize comparisons using data from overlapping years.

Data-Mining Goals

Goal 1: Vancouver's Overall Crime Rate vs. U.S. Cities

- Compute the **crime rate per capita** for each city.
- Use population data to normalize raw crime counts.

Goal 2: Crime Severity Assessment

- Assign severity weights to crime types (e.g., 1 for property crimes, 5 for violent crimes).

Goal 3: Seasonal and Time-of-Day Crime Patterns

- Extract **date and time fields** to analyze seasonal trends (e.g., summer vs. winter).
- Group data by hours of the day to identify crime spikes (e.g., night vs. daytime).

Terminology

- **Crime Rate:** Number of reported crimes per 10 000 residents annually.
- **Severity Index:** A weighted measure of crime seriousness.
- **Temporal Trends:** Changes in crime frequency by season or time of day.

Costs and Benefits

Costs

- Time for data cleaning, integration, and analysis.
- Computational resources for large-scale data processing.

Benefits

- Travel Safety Insights
    - By comparing crime rates, the analysis helps determine which cities are the safest to travel to.
- Public Awareness
    - The analysis could inform residents about crime patterns, empowering them to make safer choices regarding their daily activities.

Conclusion

This analysis provides critical insights into urban safety. Comparing Vancouver to Portland, LA, and Chicago helps understand crime disparities and severity. Seasonal and temporal patterns can inform where to not go at a given time of day. Success will be measured by the clarity, comparability, and policy relevance of the insights derived from this analysis.

Task 3. Data understanding

Gathering Data

Outline Data Requirements:
To meet the goals of the project, the following data is required:

- **Crime Incident Details:** Including crime type, timestamp, location, and severity.
- **Population Data:** Population data for each city will be obtained from separate, official sources (e.g., census data) to normalize crime counts into crime rates per capita.
- **Geographical Context:** To compare trends across Vancouver, Los Angeles, Chicago, and Portland.
- **Temporal Information:** Date and time data to analyze seasonal and time-of-day patterns.

Verify Data Availability:
The datasets have been sourced from Kaggle and include the following:

1. **Los Angeles and Chicago Crime Data (1.98 GB):** Comprehensive records covering various crime types with timestamps and locations.
2. **Portland Crime Data (84.64 MB):** Similar structure to the LA and Chicago dataset but limited in size.
3. **Vancouver Crime Data (82.12 MB):** Includes crime records across all neighborhoods and years.

All datasets appear to have the required fields; however, timestamps need verification for consistency and completeness.

Define Selection Criteria:

- **City-Level Comparison:** Focus only on records that match the chosen cities (Vancouver, Portland, Los Angeles, Chicago).
- **Time Period:** Use overlapping years for comparison to ensure uniformity. For example, if one dataset covers 2010–2020, restrict others to the same time frame.
- **Data Completeness:** Include only records with valid crime types, timestamps, and locations.

Describing Data

The datasets contain the following fields:

- **Crime Type:** Categorizes incidents into theft, assault, vandalism, etc.
- **Timestamp:** Date and time of the incident, required for temporal analysis.
- **Location:** Geographical coordinates or neighborhood identifiers for spatial analysis.
- **Crime Severity:** Recorded directly or inferred based on the crime type.

Key Characteristics:

- The **Los Angeles and Chicago dataset** is the largest, with over a million records, providing a robust sample for analysis.
- The **Portland and Vancouver datasets** are smaller but appear to offer sufficient coverage for meaningful comparison.
- Fields for timestamps and crime types are consistent across datasets, though some variation in crime categories (e.g., terminology differences) may require mapping to a standard schema.

Exploring Data

Initial exploration focuses on understanding distributions and identifying anomalies:

- **Crime Rates:** Calculate basic statistics (mean, median, standard deviation) of crime counts across years to assess trends.
- **Crime Severity:** Use weighted indices to understand distributions of minor vs. severe crimes.
- **Temporal Patterns:** Plot crime incidents by season and time of day to identify peaks and trends.
- **Location Data:** Map crimes by neighborhood or city regions to visualize spatial hotspots.

Preliminary findings include:

- **Consistent Seasonal Trends:** Summer months show a noticeable increase in crime rates across most cities.
- **Time-of-Day Patterns:** Crimes frequently spike during late evening and early night hours.

Verifying Data Quality

Completeness:

- All datasets contain the necessary fields, but some records have missing timestamps or crime types. These entries will be excluded or imputed during data preparation.

Consistency:

- Crime categories vary slightly between datasets (e.g., "Theft" vs. "Larceny"). A standard schema will be developed to map these categories uniformly.

Accuracy:

- Validate location fields (e.g., latitude and longitude) to ensure they align with city boundaries.

Data Quality Issues:

Inconsistent Reporting Standards: Some datasets group minor offenses differently.

- **Resolution:** Develop a unified categorization system.

Task 4. Project plan

Project Plan and Tasks

1. **Data Collection and Preparation (12 hours total):**
    - **Hours per Member:** Andreas (6 hours), Richard (6 hours)
    - **Description:** Collect datasets, verify availability, and preprocess data (cleaning, integration, normalization). This includes standardizing crime categories and converting timestamps.
    - **Tools/Methods:** Python (pandas, NumPy), Jupyter Notebook.
2. **Exploratory Data Analysis (EDA) (10 hours total):**
    - **Hours per Member:** Andreas (5 hours), Richard (5 hours)
    - **Description:** Perform initial data exploration, analyze distributions, and identify trends. Generate visualizations for seasonal and time-of-day patterns.
    - **Tools/Methods:** Matplotlib, Seaborn, descriptive statistics in Python.
3. **Crime Severity Index Development (8 hours total):**
    - **Hours per Member:** Andreas (4 hours), Richard (4 hours)
    - **Description:** Create a weighted severity index for crime types and compute indices for each city. Validate and interpret results.
    - **Tools/Methods:** Python (custom weighting functions).
4. **Modeling and Pattern Analysis (14 hours total):**
    - **Hours per Member:** Andreas (7 hours), Richard (7 hours)
    - **Description:** Use machine learning models to analyze crime trends and predict seasonal or temporal patterns.
    - **Tools/Methods:** scikit-learn for regression models, Python for data transformations.
5. **Visualization and Reporting (16 hours total):**
    - **Hours per Member:** Andreas (8 hours), Richard (8 hours)
    - **Description:** Develop visualizations to communicate findings clearly and compile the final report. Prepare a presentation and poster for the project showcase.
    - **Tools/Methods:** Matplotlib, Canva (for poster), Google Docs for report writing.

Methods and Tools

- **Data Preprocessing and Analysis:** Python (pandas, NumPy).
- **Visualization:** Matplotlib, Seaborn, Canva.
- **Modeling:** scikit-learn for trend analysis.
- **Documentation:** Google Docs for reports, GitHub for code repository.

Comments

- **Balanced Contributions:** Each team member will contribute equally across all tasks, ensuring both gain experience with all aspects of the project.
- **Collaboration:** Frequent meetings to synchronize progress and ensure consistent methodology.
- **Presentation:** Special attention will be given to the clarity and visual appeal of the final deliverables for the poster session.

Data sets:

Dataset 1 (1.98 GB): Datasets of crimes committed in US cities (LA and Chicago)
https://www.kaggle.com/datasets/middlehigh/los-angeles-crime-data-from-2000?select=Chicage+Crime+Data.csv
Dataset 2 (84.64 MB): Datasets of crimes committed in Portland [2]
https://www.kaggle.com/datasets/michaellindsay/portland-crime
Dataset 3 (82.12 MB): Dataset of crimes committed in Vancouver[3]
https://www.kaggle.com/datasets/tcashion/vancouver-bc-crime-dataset?select=crimedata_csv_AllNeighbourhoods_AllYears.csv