US car crash analysis

By: Loann Rio, Lucas Marais

Task 2: Developing a Business Understanding

---

1. Identifying Business Goals

Background
Car crashes in united states have a large societal and economical impact, including loss of life, injuries and important financial costs. These crashes are due to a lot of different factors outside of the driver himself, such as the road infrastructure, the population density and the weather conditions. By analyzing large-scale crash and weather conditions datasets, the goal of this project is to uncover elements that can help to improve public safety and urban planning by reducing potential avoidable danger.

Business Goals

1- **Identifying High-Risk Zones**
This goal focuses on detecting geographic regions where car crashes are most frequent and severe. The insights will help policymakers, transportation authorities, and urban planners prioritize safety interventions such as improved road design, improved signage, or targeted enforcement of traffic laws.

- **Output**: A detailed heatmap or geographic visualization of high-risk zones.
- **Applications**: Allocate resources effectively, design safer road infrastructure, and implement area-specific traffic management strategies.

2- **Understanding Weather Impact on Accidents**
This goal involves analyzing the role of weather conditions—such as rain, snow, fog, and extreme temperatures—in influencing crash frequency and severity. The objective is to identify patterns that make certain weather conditions more dangerous for drivers, helping authorities devise better safety measures during adverse weather events.

- **Output**: A report quantifying the relationship between weather conditions and crash probability, supported by predictive models.
- **Applications**: improve weather-specific traffic advisories, inform vehicle safety features, and guide public awareness campaigns.

Business Success Criteria

The success of the project will be determined by its ability to provide actionable insights and reliable outputs that can be used by stakeholders to improve road safety. The specific criteria for evaluating success are as follows:

The success of the project can be determined by its ability to give reliable information about the impact of certain weather conditions that can be used to improve road safety. To quantify the success of the project, we choosed a set of different criteria:

1. Accuracy and Reliability of Outputs
    - High-risk zones are identified with at least 80% accuracy in spatial clustering analysis.
    - Predictive models correlate weather conditions with crash risks, achieving a minimum 70% accuracy in prediction metrics.
2. Practicality and Usability of Findings
    - The information are presented in a format that is easily interpretable by policymakers, and urban planners.
    - Outputs such as GIS heatmaps, visualizations, and reports should directly inform decision-making processes, such as prioritizing areas for intervention.
3. Stakeholder Impact
    - The project need to be able to give help determine recommendation that can be applied to improve road safety by implementing new measures such as targeted investment on infrastructure and improved weather recommendations.
    - Stakeholders report a positive impact on their ability to make data-driven decisions within six months of the project's conclusion.
4. Comprehensive Coverage
    - The project analyzes a representative subset of crash and weather data, ensuring findings are statistically valid and geographically comprehensive.
    - Missing data issues are addressed effectively, ensuring outputs are robust and reliable.

By meeting these criteria, the project will ensure that its outputs are not only technically sound but also practically valuable, contributing to measurable improvements in traffic safety outcomes.

If the project meet these criteria, we can assert that the output will not be only technically interesting but also useful to contribute to improvement in public safety in traffic giving it a practical value.

---

2. Assessing the Situation

Inventory of Resources

- Datasets:
    - Crash data (3GB) from GitHub [Dataset 1].
    - Weather data (1GB) from Kaggle [Dataset 2].

- Tools: Python, libraries (Pandas, GeoPandas, Matplotlib, Scikit-learn), GIS tools like OpenStreetMap.

Requirements, Assumptions, and Constraints

- Requirements: Robust computational resources to process large datasets.
- Assumptions: Data fields from both datasets are temporally and spatially aligned.
- Constraints: Time limitations for analysis, potential gaps in dataset completeness.

Risks and Contingencies

- Risks: Missing data or incomplete fields; discrepancies between crash and weather data.
- Contingencies: Use supplementary datasets if necessary, or focus analyses on subsets of reliable data.

Terminology

- High-Risk Zones: Geographic areas with a higher probability of car crashes.
- Weather Impact: The effect of weather events (e.g., rain, snow) on accident probability.

Costs and Benefits

- Costs: Time investment, computational resources, and potential costs for datasets.
- Benefits: useful information to improve public safety, help urban planners, and reduce accident-related costs.

---

3. Defining Data-Mining Goals

Data-Mining Goals

- Identify geographic high-risk zones using clustering algorithms and GIS techniques.
- Develop predictive models to quantify and understand the impact of weather conditions on crash frequency.

Data-Mining Success Criteria

- Clustering models achieving at least 80% accuracy in identifying high-risk zones.
- Predictive models with at least 70% accuracy in correlating weather events with crash risks.

Task 3: Data Understanding

---

1. Gathering Data

Outline Data Requirements

- Fields: Crash location (latitude/longitude), severity, time, vehicle type, weather conditions, road type.

- Timeframe: Last five years.

Verify Data Availability

- Dataset 1: Car crash data available (3GB) via GitHub repository.
- Dataset 2: Weather event data available (1GB) from Kaggle.

Define Selection Criteria

- Data from regions with comprehensive crash and weather records.

---

2. Describing Data

Crash Dataset (Dataset 1)

- Contains details on crash location, severity, time, and other factors like vehicle type and road characteristics.
- Approximately 3GB of data spanning multiple states and years.

Weather Dataset (Dataset 2)

- Includes event type (rain, snow, etc.), intensity, time, and geolocation.
- Approximately 1GB, providing coverage of weather events in the US.

---

3. Exploring Data

Initial Observations from Dataset 1:

- Geographic Distribution: Data is unevenly distributed, with higher density in urban areas.
- Severity Trends: A higher proportion of minor crashes, with severe accidents concentrated near highways and intersections.
- Temporal Patterns: Peaks in crashes during holidays and rush hours.

Initial Observations from Dataset 2:

- Weather Event Types: Rain and snow are predominant. Visibility and wind conditions vary widely.
- Temporal Alignment: Weather events align with seasons and geographic regions (e.g., hurricanes in coastal areas, snow in northern regions).

Preliminary Visualizations :

- Heatmaps of crash locations.
- Bar charts for crash severity versus weather conditions.
- Time-series plots to correlate accident frequency with weather patterns.

The data is presented in the following format:

```
Accidents Dataset Info
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7728394 entries, 0 to 7728393
Data columns (total 46 columns):
 #   Column                 Dtype
---  ------                 -----
 0   ID                     object
 1   Source                 object
 2   Severity               int64
 3   Start_Time             object
 4   End_Time               object
 5   Start_Lat              float64
 6   Start_Lng              float64
 7   End_Lat                float64
 8   End_Lng                float64
 9   Distance(mi)           float64
 10  Description            object
 11  Street                 object
 12  City                   object
 13  County                 object
 14  State                  object
 15  Zipcode                object
 16  Country                object
 17  Timezone               object
 18  Airport_Code           object
 19  Weather_Timestamp      object
 20  Temperature(F)         float64
 21  Wind_Chill(F)          float64
 22  Humidity(%)            float64
 23  Pressure(in)           float64
 24  Visibility(mi)         float64
 25  Wind_Direction         object
 26  Wind_Speed(mph)        float64
 27  Precipitation(in)      float64
 28  Weather_Condition      object
 29  Amenity                bool
 30  Bump                   bool
 31  Crossing               bool
 32  Give_Way               bool
 33  Junction               bool
 34  No_Exit                bool
 35  Railway                bool
 36  Roundabout             bool
 37  Station                bool
 38  Stop                   bool
 39  Traffic_Calming        bool
 40  Traffic_Signal         bool
 41  Turning_Loop           bool
 42  Sunrise_Sunset         object
 43  Civil_Twilight         object
 44  Nautical_Twilight      object
 45  Astronomical_Twilight  object
dtypes: bool(13), float64(12), int64(1), object(20)
memory usage: 2.0+ GB
None

Weather Events Dataset Info
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8627181 entries, 0 to 8627180
Data columns (total 14 columns):
 #   Column             Dtype
---  ------             -----
 0   EventId            object
 1   Type               object
 2   Severity           object
 3   StartTime(UTC)     object
 4   EndTime(UTC)       object
 5   Precipitation(in)  float64
 6   TimeZone           object
```

```
 7   AirportCode       object
 8   LocationLat       float64
 9   LocationLng       float64
 10  City              object
 11  County            object
 12  State             object
 13  ZipCode           float64
dtypes: float64(4), object(10)
memory usage: 921.5+ MB
None
```

---

4. Verifying Data Quality

4.1 Completeness

- Check for missing values in critical fields (e.g., latitude, longitude, severity).
- Ensure timestamps are complete and standardized across datasets.

```
Missing values in Accidents dataset:
ID                          0
Source                      0
Severity                    0
Start_Time                  0
End_Time                    0
Start_Lat                   0
Start_Lng                   0
End_Lat               3402762
End_Lng               3402762
Distance(mi)                0
Description                 5
Street                  10869
City                      253
County                      0
State                       0
Zipcode                  1915
Country                     0
Timezone                 7808
Airport_Code            22635
Weather_Timestamp      120228
Temperature(F)         163853
Wind_Chill(F)         1999019
Humidity(%)            174144
Pressure(in)           140679
Visibility(mi)         177098
Wind_Direction         175206
Wind_Speed(mph)        571233
Precipitation(in)     2203586
Weather_Condition      173459
Amenity                     0
Bump                        0
Crossing                    0
Give_Way                    0
Junction                    0
No_Exit                     0
Railway                     0
Roundabout                  0
Station                     0
Stop                        0
Traffic_Calming             0
Traffic_Signal              0
Turning_Loop                0
Sunrise_Sunset          23246
Civil_Twilight          23246
Nautical_Twilight       23246
Astronomical_Twilight   23246
dtype: int64

Missing values in Weather Events dataset:
```

```
EventId             0
Type                0
Severity            0
StartTime(UTC)      0
EndTime(UTC)        0
Precipitation(in)   0
TimeZone            0
AirportCode         0
LocationLat         0
LocationLng         0
City            16912
County              0
State               0
ZipCode         69199
```

4.2 Accuracy

- Compare random samples of crash locations with OpenStreetMap geospatial data for consistency.
- Validate weather data accuracy using external benchmarks if available.

4.3 Consistency

- Ensure uniform formats for time and location fields across datasets.
- Remove duplicate records, particularly in crash data.

4.4 Relevance

- Filter out irrelevant records (e.g., crashes without injuries or property damage).
- Exclude weather events unrelated to crashes (e.g., distant hurricanes).

Task 4: Planning Your Project

---

1. Project Plan

Planned Tasks and Time Allocation

1. Data Collection and Preprocessing (10 hours)
   - Merge datasets and clean for consistency and completeness.
   - Address missing values and normalize fields for compatibility.
2. Exploratory Data Analysis (5 hours)
   - Conduct statistical and visual analyses to uncover patterns.
   - Use heatmaps and plots to visualize crash distributions and weather impacts.
3. Model Development (20 hours)
   - Develop clustering models for identifying high-risk zones.
   - Build regression and classification models for analyzing weather impact.

4. Visualization and Reporting (10 hours)
    - Create visual outputs such as GIS heatmaps and correlation charts.
    - Compile findings into a clear report.
5. Final Review and Deployment (10 hours)
    - Validate findings for accuracy and usability.

---

2. Methods and Tools

- Methods: Clustering algorithms (e.g., K-Means), regression models, correlation analysis.
- Tools: Python, GIS tools (OpenStreetMap), and data visualization libraries (Matplotlib, Seaborn).
