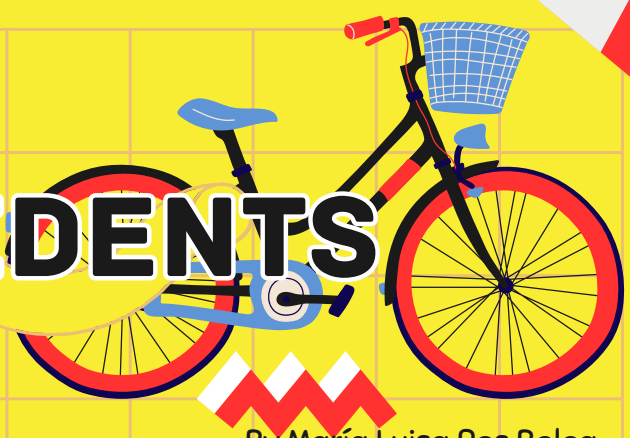
ANALYSIS OF BICYCLE ACCIDENTS IN MADRID



By María Luisa Ros Bolea

PROJECT OBJETIVES

The primary objective of this project was to analyze bicycle accidents in Madrid during the year 2025. The goal was to identify patterns, relationships, and risk factors related to accidents involving cyclists, considering variables such as gender, age, accident type, district, weather conditions, alcohol and drug influence, vehicle type, and injury severity.

By understanding these relationships, the analysis aims to provide actionable insights for urban safety planning, preventive measures, and awareness campaigns targeted at the most at-risk groups.

DATASET DESCRPTION

The dataset consisted of 827 records with 19 columns, including identifiers, dates, locations, accident types, vehicle types, victim characteristics, injury severity, and alcohol/drug test results.

Key steps in data preparation:

1. Importing and cleaning the dataset:

- Ensured that missing values were identified.
- Corrected inconsistent entries in categorical variables.
- Checked data types for numeric and categorical columns.

2.Columns of interest for analysis:

- gender
- age_range
- accident_type
- district
- weather_conditions
- vehicle_type
- injury_severity
- alcohol_positive
- drug_positive

3.Handling missing values:

- Some columns had missing values (lesivity and drug_positive). These were either filtered out for specific analyses or treated as separate categories to avoid bias.

DESCRIPTIVE ANALYTICS

The first step in understanding the data involved generating basic statistics and distributions for both numeric and categorical variables.

Key findings:

- **Numeric analysis:**
  a. District codes ranged from 1 to 21.
  b. Coordinates confirmed the geospatial consistency of accident locations.
  c. Injury codes ranged from minor to severe injuries (1–14), with a majority in the minor/moderate range.
- **Categorical analysis:**
  a. Gender distribution: majority male victims.
  b. Age distribution: young adults (18–30) accounted for a significant portion of accidents.
  c. Accident types: collisions were the most frequent, followed by falls and other incidents.
  d. Vehicle types involved: bicycles and light vehicles dominated the dataset.
  e. Weather conditions: most accidents occurred under normal conditions, with fewer in adverse weather.

VISUAL ANALYSIS

Visualization was a critical part of the analysis to identify patterns and facilitate interpretation.

Charts and graphs:

Bar charts and histograms:

- Distribution of accidents by gender, age range, and accident type.
- Distribution of accidents by district to identify high-risk areas.

Heatmaps of cross-tabulations:

- Gender vs Accident Type
- Age Range vs Injury Severity
- District vs Accident Type
- Age vs Alcohol/Drug Positivity
- Vehicle Type vs Accident Type

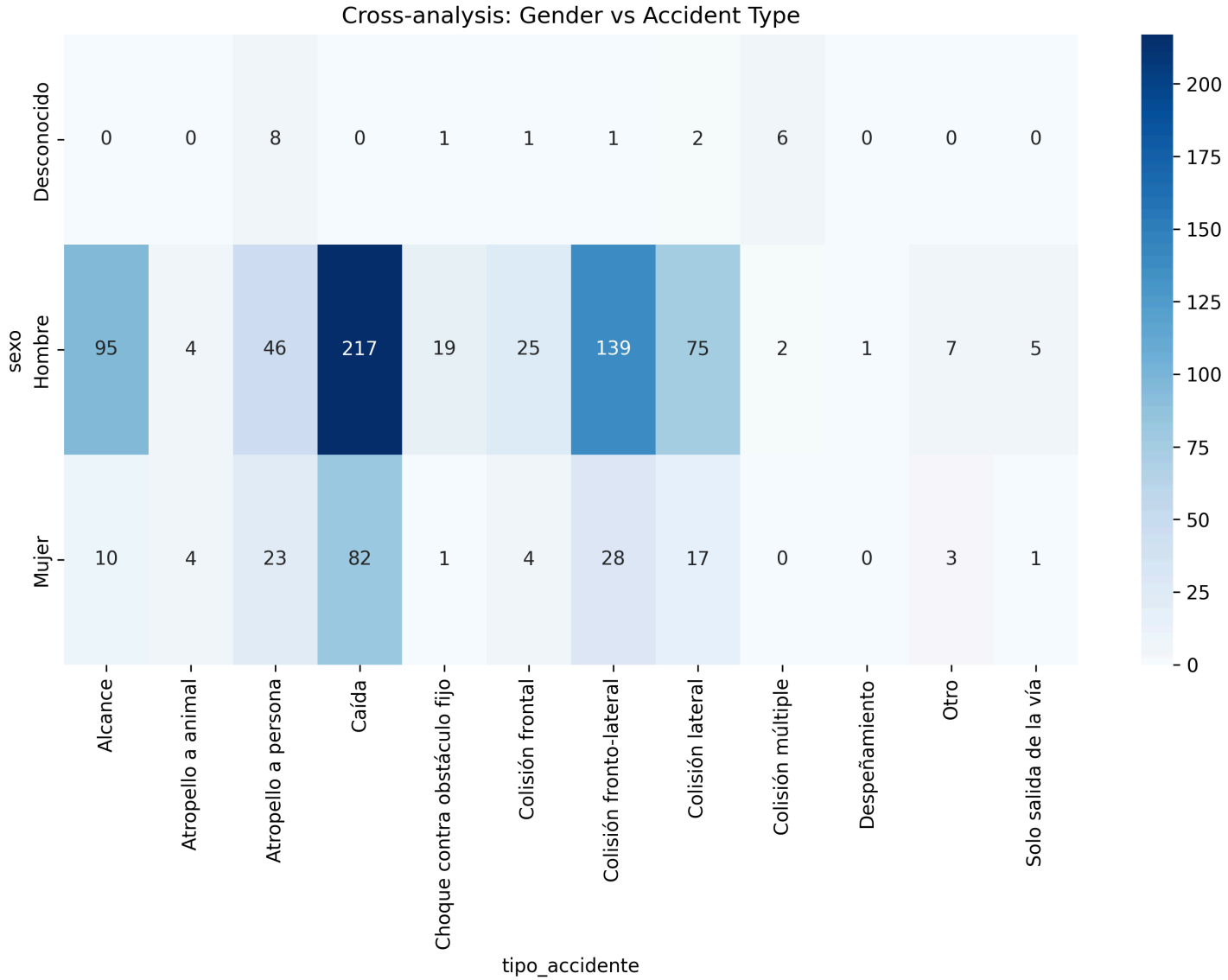
All visualizations were created using a blue color palette for consistency and clarity. Each chart was exported using plt.savefig() for further use in reports and presentations.

CROSS TABULATION ANALYSIS

To uncover relationships between categorical variables, we used cross-tabulations and heatmaps:

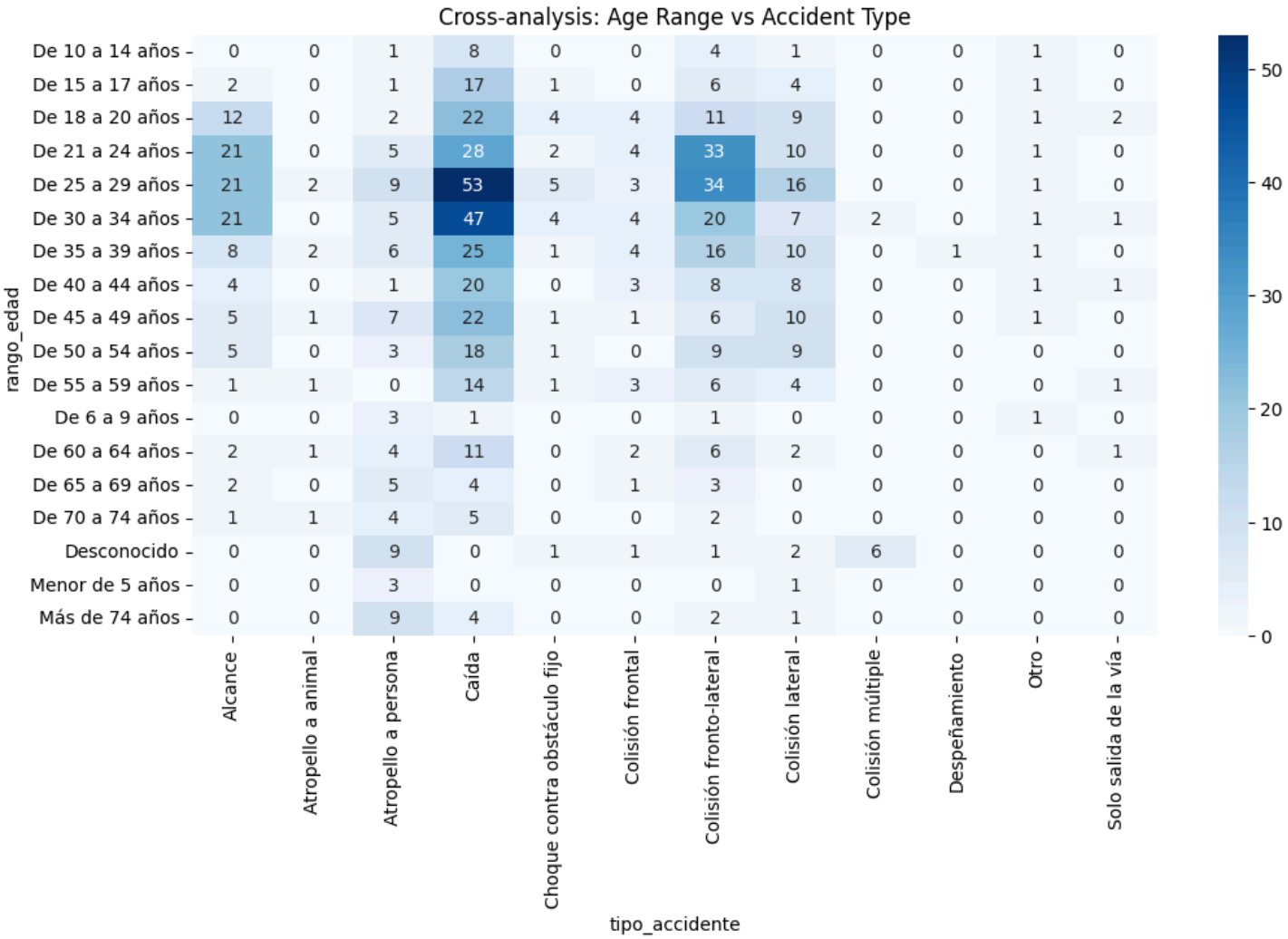
1. Gender vs Accident Type

Men show higher counts across almost all accident types.



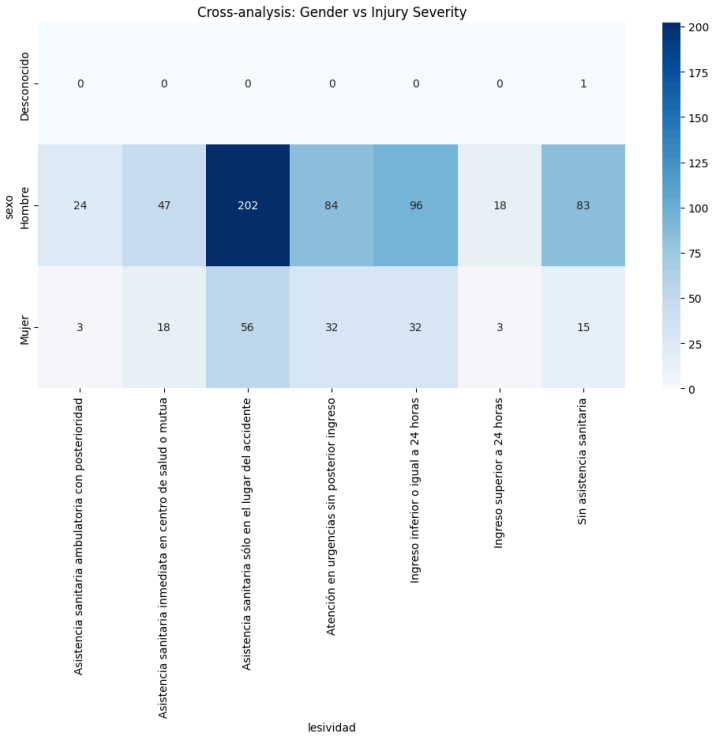
1. Gender vs Injury Severity

Most men sustain moderate injuries; women are less frequent but some severe injuries occur.



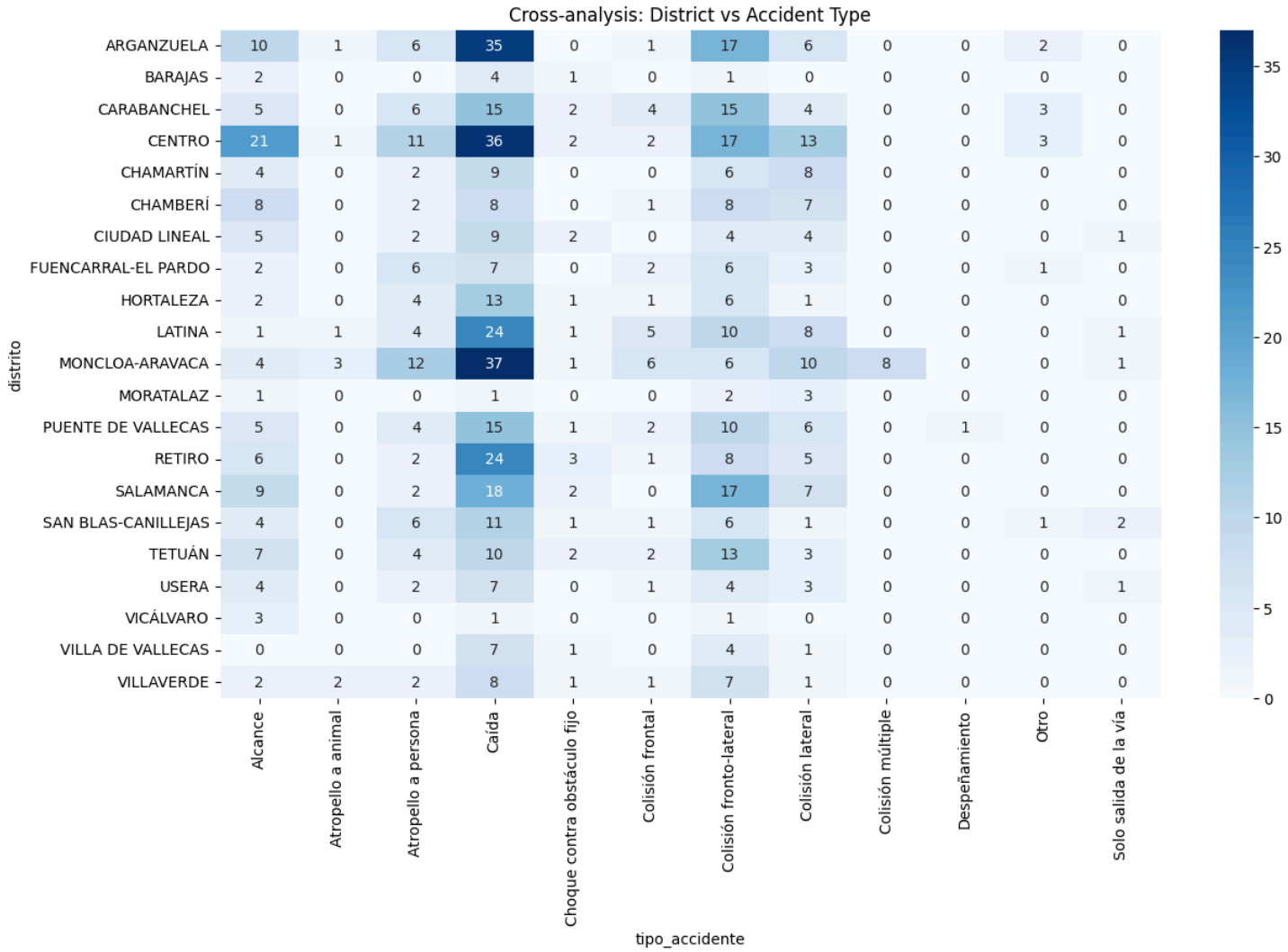
Age Range vs Accident Type

Young adults (18–30) are more likely to experience falls and collisions. Older age ranges are less frequent but more severe accident.



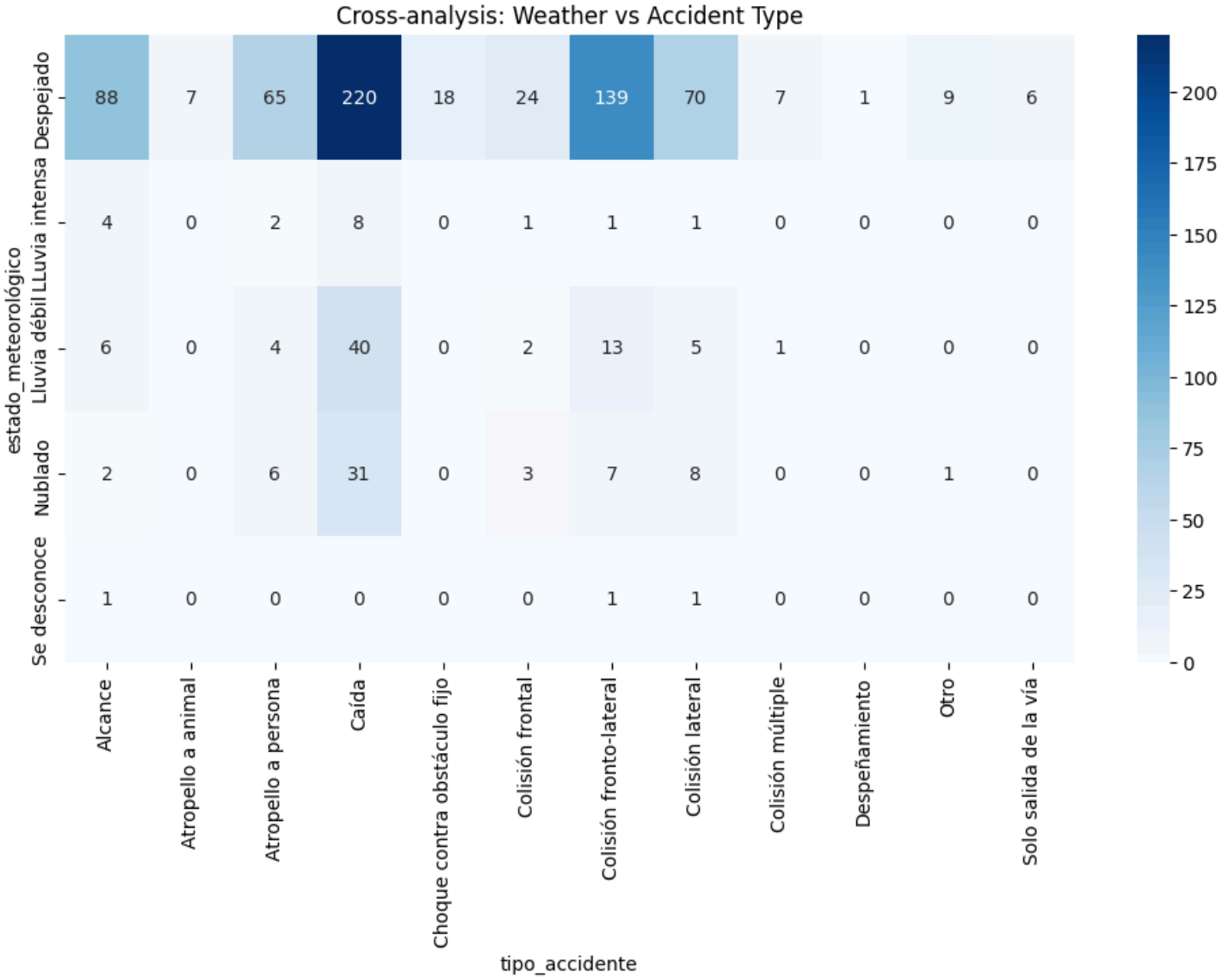
District vs Accident Type

Central districts exhibit higher collision and fall counts.



Weather vs Accident Type

Adverse conditions like rain slightly increase accident occurrences.



PREDICTIVE MODELING

We implemented a Decision Tree Classifier to explore whether accident characteristics can predict injury severity (lesividad).

Modeling Steps

1. Selected features: sexo, rango_edad, distrito, tipo_accidente, estado_meteorológico.
2. Removed rows with missing target values.
3. Converted categorical variables to numeric using LabelEncoder.
4. Split data into training (80%) and testing (20%) sets.
5. Trained a decision tree (max_depth=5) to avoid overfitting.

Evaluation

- Accuracy: ~38%
- Confusion matrix: Shows that the model predicts common classes better but struggles with rare categories.
- Insights: Despite low accuracy, the model illustrates how machine learning can assist in understanding risk factors and predicting potential accident severity.

KEY FINDINGS

- Most bicycle accidents occur during commuting hours (8–9h, 18–20h).
- Male cyclists are more frequently involved than females.
- Certain districts have higher accident concentrations, suggesting localized risk areas.
- Age is a critical factor: older cyclists tend to sustain more severe injuries.
- Weather conditions, particularly rain, have a small but noticeable impact on accident frequency.
- Cross-tab heatmaps clearly show correlations between gender, age, district, accident type, and injury severity.

RECOMMENDATIONS

- Develop targeted awareness campaigns for high-risk groups (young male cyclists).
- Enhance traffic safety measures in districts with higher accident density.
- Encourage helmet and protective gear usage among older cyclists.
- Use predictive models to identify potential high-risk scenarios and guide urban safety planning.

CONCLUSIONS

1. This project provides a comprehensive analysis of bicycle accidents in Madrid for 2025, combining:
2. Descriptive statistics
3. Cross-tab analyses with heatmaps
4. Predictive modeling for injury severity
5. The findings highlight the importance of demographics, geography, and environment in accident occurrence and severity. Cross-analytics and machine learning approaches demonstrate actionable insights for public safety strategies and urban planning interventions.

Whole proyect:

GitHub - malurosbolea-ux/Colab-bicycle-accidents-in-madrid-repo: Data analysis project on bicycle accid...