Summary

Our mission is to conduct an exploratory analysis of automobile accidents in the city of Chicago using a comprehensive dataset consisting of six main components: automobile accidents, individuals involved, vehicles, police district boundaries, central business district boundaries, and city street boundaries. Using the power of hidden data, we aim to gain valuable insights into the patterns and factors influencing automobile accidents within the city.

Through this analysis, we seek to highlight the importance of public safety, traffic planning and regulation. By understanding the complex network of car accidents, we can identify high-risk areas and common factors that lead to accidents. With this knowledge, efforts can be directed towards improving road safety measures, infrastructure planning and traffic management to create a safer and more harmonious transportation environment.

Questions

1. **Data cleaning and integration**
    a. **Filling gaps**
        - We have data from different sources and need to combine them to get a more complete picture of car accidents in the city of Chicago. While these datasets have a lot of columns (more than 20 columns per dataset), be aware that they may not contain enough information. Identify the columns you need and explain why you are omitting the rest of the columns in each table in the database.
        - Before starting the analysis process, you must clean the data set, i.e. you must solve the problems of empty values, extreme values, and contradictions in a way that you see fit.

    b. **From raw to polished!**
        i. Extract the year in which the traffic accident occurred and put it in a new column similar to the month and hour column.

ii. Use the table of individuals involved in the accident and calculate the number of passengers in each vehicle (other than the driver) and then calculate the average number of passengers and the average age of the passenger in each traffic accident in the database.
iii. Use the vehicle table to calculate the number of vehicles involved in a traffic accident.
iv. Take advantage of the vehicle table and determine the vehicle age category (very old, old, new) by taking advantage of the manufacturing date column and the accident date by choosing a certain threshold that is appropriate from your point of view for classification, then create a multi-category for the age categories of the vehicles involved in the traffic accident in the form of a comma separated list where the list contains only the unique categories regardless of the repetition of each category.
v. Perform a binning operation the list of vehicle age classes as follows (Old, New, Mixed) and convert it to a conventional nominal attribute (Categorical).
vi. Group the geographic locations of traffic accidents into geographic segments using the geographic hashing technique and set a resolution appropriate from your point of view for the number of bits used in the hashing process.
vii. Perform a segmentation process for the geographical sectors you obtained in the geographical hashing process into 3 categories according to the number of traffic accidents in each sector (red, yellow, gray), specifying the boundaries of each category in a way you see fit.
viii. Calculate the length of each street in Chicago in kilometers. Then, discretize it into three categories (short, medium, long), specifying the boundaries of each category in a way that you see fit. Then, combine the result you obtained with the traffic accident table.

ix. Use the geographic polygon that defines the central business district within the city to calculate the distance of each incident from this district in kilometers, then perform a segmentation process into three categories (near, medium, far).
x. Perform a Binarization of the Distance from Business Zone column to determine whether the incident was within or outside the zone.
xi. Use the county boundaries table to determine the county in which each traffic accident occurred.
xii. Create any feature that might help you in the exploration or integration process.

Tip: Perform a projection of the geographic polygon to a projection of equal metric area - homogeneous - (EPSG:6933) before calculating the distance of the geographic point from that polygon. Remember that the Earth is not perfectly spherical and therefore the result will vary depending on the projection.

Reminder: The process of feature engineering may bring new problems and you should solve them and pay attention to the different cases and what is the real benefit of each feature so as not to cause loss of information.

2. Exploration and analysis

a. Study the causes of traffic accidents

i. Draw a bar chart of the number of accidents by primary cause as estimated by the police officer (PRIM_CONTRIBUTORY_CAUSE).
ii. Draw a bar chart of the number of accidents by lighting condition (LIGHTIING_CONDITION) and the damage category estimated by the police officer (Damage).
iii. Draw a bar chart of the number of accidents according to the severity rating (CRASH_TYPE) and the first collision type (FIRST_CRASH_TYPE).

iv. Draw a bar chart of the number of accidents by road type (TRAFFICWAY_TYPE) and lighting conditions (LIGHTIING_CONDITION).
v. Try to explain the results using your domain knowledge.

b. Study the time of traffic accidents

i. Draw a bar chart for the number of traffic accidents by hour (CRASH_HOUR) and the category of damage estimated by the police officer (Damage).
ii. Draw a line chart for the total number of accidents per month of each year (TOTAL COUNT PER MONTH PER YEAR).
iii. Draw a box plot to show the distribution of monthly accidents over a year.
iv. Plot a Sunburst chart of the number of accidents by day of the week (CRASH_DAY_OF_WEEK) and the damage category estimated by the police officer (DAMAG).
v. Try to explain the results using your domain knowledge.

c. Spatial analysis of traffic accidents

i. Draw a bar chart of the number of traffic accidents in each geographic sector.
ii. Draw a box plot of the distance from the central business district by damage category (estimated by the police officer).
iii. Draw a box plot of the length of the street within the city.
iv. Draw a bar chart of the number of traffic accidents by street length category.
v. Draw a bar chart of the number of traffic accidents by distance category from the central business district.
vi. Is there a relationship between the street length category and the distance category from the central business district (Chi Square Analysis)?
vii. Try to explain the results using your domain knowledge.

d. Passenger case study

i. Draw a Scatter diagram where each point represents (driver's age, driver's gender, age of the vehicle on the date of the traffic accident).
ii. Is there a relationship between the age group of the car and the age group of the driver (Chi Square Analysis)?
iii. Draw a box plot for the average age of passengers excluding the driver.
iv. Draw a box plot for the number of passengers excluding the driver.
v. Draw a bar chart for the age category of the vehicle.
vi. Try to explain the results using your domain knowledge.

e. Extreme cases study

i. Set the group of extreme or strange traffic accidents in a way that you see fit.
ii. Try to explore the reasons for the extremism of these cases, expressing your opinion.

Tip: You can use Cramer's V to measure the strength of the association if any.

Note: Any diagram that is not explained by the student in the notebook will be ignored, and this will result in the deduction of the mark for requests related to the diagram, in whole or in part, as the teacher deems appropriate.

3. Research skill

Find a research paper that studies how to handle disguised missing values. Understand the method and summarize it in no more than two pages, taking into account that the research is as recent as possible (from 2020). Write the following in the summary:

- Name of the publishing journal
- Explanation of the solution method
- Data sets used for the test and test results
- Disadvantages or limitations of the method

4. Secret mission!

Knowledge mining may seem daunting at first, but remember that every piece of data holds valuable insights waiting to be uncovered. With your analytical skills and determination, you can dive deep into the data, explore its intricacies, and extract meaningful patterns and facts. While it is important to adhere to the requirements of the homework, do not let them limit your journey of discovery. Allow yourself to think outside the box and consider different angles and perspectives.

You can use any table from the database as per your need, draw any chart or calculate any statistic to support your conclusions, and remember that mining is a dynamic and iterative process, so make sure to write clean and readable code that is commonly used and try to organize the notebook as much as possible.

Check out the pipes function in the Pandas library and use as many functions in the library as possible instead of reinventing the wheel!

Stay curious, be creative, and let data guide your journey. Storytelling is one of the most in-demand skills when combined with data science! The way you present it to your audience is much more important than the results.