Friedrich-Alexander-Universität
Erlangen-Nürnberg

METHODS OF ADVANCED DATA ENGINEERING

Project:- Road Accident & Bicycle Traffic Trends in Cologne

Presented By:

Ankit Joshi
23123817
qa97lela

Introduction

Data Sources

Data Engineering Pipeline

Exploratory Data Analysis

Results & Future Work

Project 1:- Analysis of Road Accidents in Cologne

Project Scope

This project embarks on an extensive analysis of traffic-related incidents that occurred within the city of Cologne throughout the calendar years of 2017 and 2018. In unraveling the stories hidden within every accident, our project focuses on Cologne's streets in 2017 and 2018, seeking insights for a safer future

Focus of the Analysis

This study dives into three main aspects: understanding the types of accidents, how lighting conditions affect incident rates, and the impact of road conditions. Additionally, we'll closely examine the unique characteristics that set apart bicycle accidents from vehicular ones.

Project Goal

This project is driven by a dedication to urban safety and preventing accidents strategically. By breaking down statistical patterns systematically, we aim to provide stakeholders with evidence-based insights. The ultimate goal is to contribute to creating a city environment marked by careful planning and improved safety standards.

Project 2:- Bicycle Traffic in Cologne: An increase or decrease in the traffic

Project Scope

This project explores Cologne's cycling trends using data from 17 counting points, spanning 2020 to 2022. The online platform makes it easy to explore individual points, providing a detailed look at bicycle traffic dynamics in the city.

Focus of the Analysis

This analysis looks at how bicycle traffic has changed over the years, using data from 2020 and 2021. We're trying to find out if cycling has increased, decreased, or stayed about the same. The online mapping tool helps us understand these changes more thoroughly.

Project Goal

Our main goal is to provide useful information about how bicycle traffic is changing. This information can help with city planning and efforts to create more sustainable and cyclist-friendly environments in Cologne. I want to share data-driven insights with decision-makers, urban planners, and the community.

Methods

Data Source

We gathered crucial datasets from reputable sources, such as the accident atlas and police accident statistics for road accidents, and automatic counting points for cycling traffic. These sources provide comprehensive insights into accident locations, types, and conditions, as well as trends in bicycle traffic over the years. The selection of reliable and diverse data sources forms the cornerstone of our analysis.

Data Engineering (Cleaning and Pipeline Creation)

This stage involves cleaning and transforming the raw data to make it suitable for exploration. Irrelevant columns are dropped, and necessary columns are renamed for clarity. Additionally, values are replaced with their actual meanings to enhance interpretability. The cleaned data is then structured into a systematic pipeline, streamlining the subsequent analysis.

Exploratory Data Analysis

With the cleaned dataset in place, the Exploratory Data Analysis (EDA) stage begins. Here, statistical and visual techniques are employed to unearth patterns, trends, and insights within the data. EDA allows us to understand the distribution of accidents, identify key factors influencing incidents, and make initial observations that will guide more in-depth analysis.

Results

The final stage focuses on presenting the analyzed data in a fitting format without interpretation. This may include creating tables, diagrams, figures, or similar visual representations that effectively convey the findings. The results stage lays the groundwork for the subsequent interpretation and discussion of the project's outcomes.

Introduction

Data Sources

Data Engineering Pipeline

Exploratory Data Analysis

Results & Future Work

Data Sources

Data Sources and Dataset Description for Project 1:

For this analysis, two datasets were employed, accessible via the following URLs:

2018 Dataset - https://offenedaten-koeln.de/sites/default/files/Unfallstatistik%20K%C3%B6ln%202018.csv

2017 Dataset - https://offenedaten-koeln.de/sites/default/files/Unfallstatistik%20K%C3%B6ln%202017.csv

The datasets contain detailed information on traffic-related incidents in Cologne for the respective years, encompassing variables such as the year of occurrence, month, hour, weekday, accident category, accident type, lighting conditions, involvement of bicycles and cars, and road conditions.

Data Sources

Data Sources and Dataset Description for Project 2:

The primary data sources for this analysis include two datasets from bicycle counting points in Cologne. The datasets were obtained from the following URLs:

2022 Dataset - https://offenedaten-koeln.de/sites/default/files/Radverkehr%20f%C3%BCr%20Offene%20Daten%20K%C3%B6ln%202022.csv

2021 Dataset - https://offenedaten-koeln.de/sites/default/files/Radverkehr%20f%C3%BCr%20Offene%20Daten%20K%C3%B6ln%202021.csv

2020 Dataset - https://offenedaten-koeln.de/sites/default/files/Fahrrad_Zaehlstellen_Koeln_2020.csv

This dataset provides a rich temporal perspective on bicycle traffic dynamics within the city. The counting points offer a comprehensive view, capturing the ebb and flow of cycling activity over the years.

Introduction

Data Sources

Data Engineering Pipeline

Exploratory Data Analysis

Results & Future Work

Data Engineering Steps

01	Combining Datasets	• Merged the datasets for 2017 and 2018 into a single dataframe for comprehensive analysis.
02	Dropping Irrelevant Columns and Renaming the Remaining Columns	• Removed columns that were deemed irrelevant for the current analysis. • Standardized column names for clarity and consistency.
03	Removing Row and Column Errors and Handling Null Values	• Character encoding errors were addressed in column names and rows by replacing characters like 'Ã¼' with 'ü', 'ÃŸ' with 'ß', and 'Ã¤' with 'ä'. • NaN values were replaced with relevant values to ensure consistency and facilitate downstream analysis.
04	Replacing Values with Actual Meaning	• Substituted coded values with meaningful labels for better interpretation.
05	Database Connection and Storage	• Established a SQLite database connection and stored the processed data frames ie. road_accidents as tables within the database for efficient retrieval and analysis.

Introduction

Data Sources

Data Engineering Pipeline

Exploratory Data Analysis

Results & Future Work

Exploratory Data Analysis for Project 1

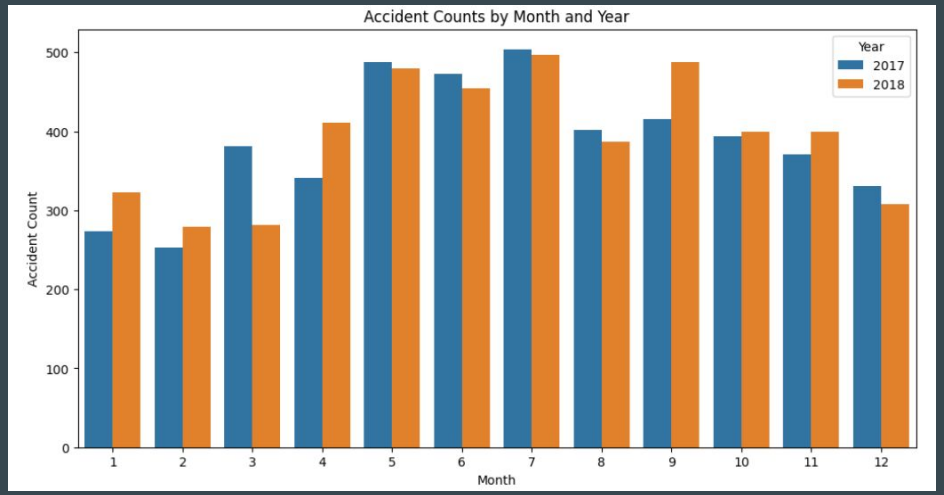
```
# Replace 'your_database_file.sqlite' with the actual name of your SQLite file
database_file = 'C:/Users/DELL/Downloads/Dataset.sqlite'

# Establish a connection to the SQLite database
connection = sqlite3.connect(database_file)

# Create a cursor object to interact with the database
cursor = connection.cursor()

#Fetch all rows from table
sql_query = 'SELECT * FROM road_accidents'
road_accidents = pd.read_sql_query(sql_query, connection)

# Remember to close the connection when you're done
connection.close()
```

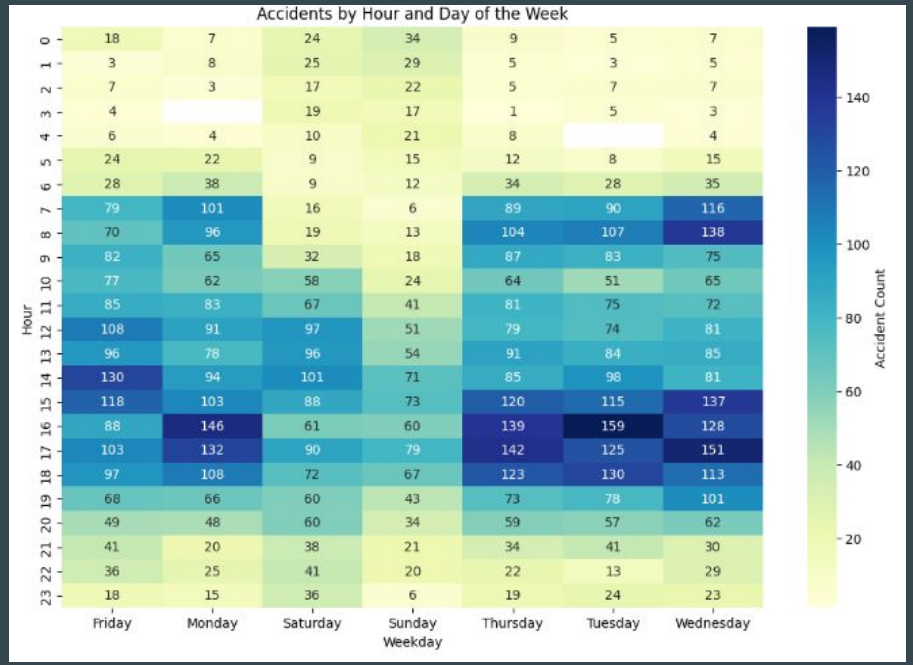
```
# Pie Chart for Accident Type
plt.figure(figsize=(6, 6))
road_accidents['Accident_Type'].value_counts().plot.pie(autopct='%1.1f%%', ylabel='')
plt.title('Distribution of Accident Type')
plt.show()
```

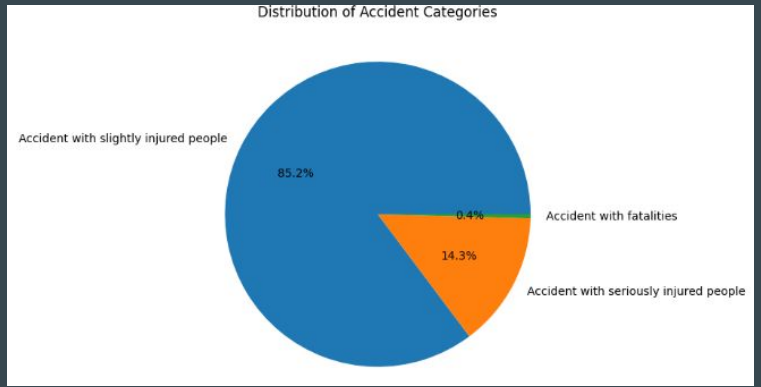
```
road_accidents['Month'] = road_accidents['Month']
road_accidents['Year'] = road_accidents['Year']

# Bar Chart for Accident Counts by Month and Year
plt.figure(figsize=(12, 6))
sns.countplot(x='Month', hue='Year', data=road_accidents)
plt.title('Accident Counts by Month and Year')
plt.xlabel('Month')
plt.ylabel('Accident Count')
plt.legend(title='Year')
plt.show()
```

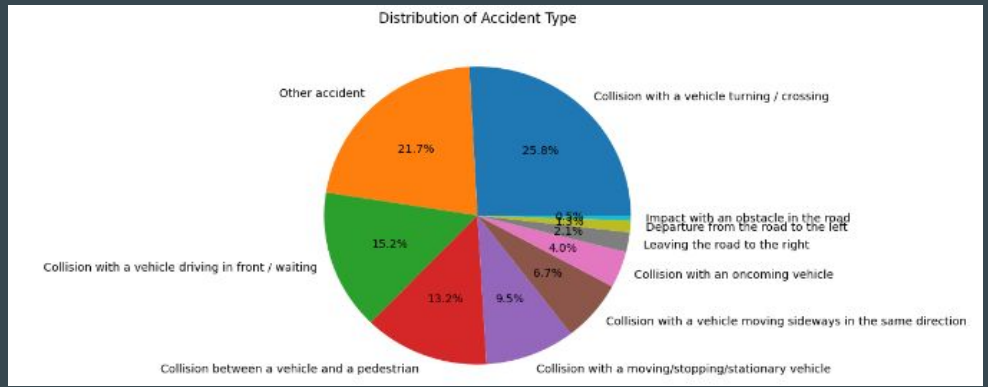
```
# Heatmap for Accidents by Hour and Day of the Week
accidents_heatmap = road_accidents.groupby(['Hour', 'Weekday']).size().reset_index(name='Accident_Count')
accidents_heatmap = accidents_heatmap.pivot('Hour', 'Weekday', 'Accident_Count')

plt.figure(figsize=(12, 8))
sns.heatmap(accidents_heatmap, cmap='YlGnBu', annot=True, fmt='g', cbar_kws={'label': 'Accident Count'})
plt.title('Accidents by Hour and Day of the Week')
plt.xlabel('Weekday')
plt.ylabel('Hour')
plt.show()
```

Exploratory Data Analysis









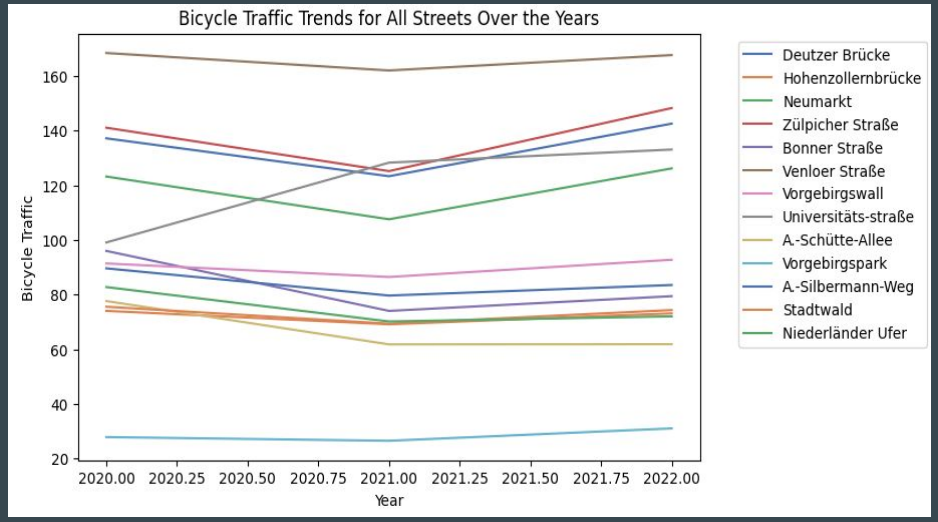
Exploratory Data Analysis for Project 2

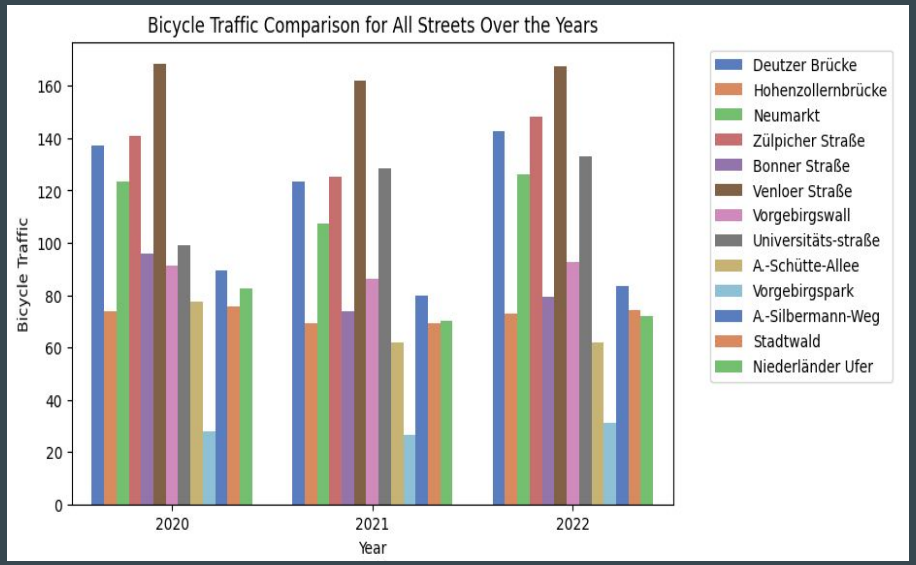
```
# Replace 'your_database_file.sqlite' with the actual name of your SQLite file
database_file = 'C:/Users/DELL/Downloads/Dataset.sqlite'

# Establish a connection to the SQLite database
connection = sqlite3.connect(database_file)

# Create a cursor object to interact with the database
cursor = connection.cursor()

#Fetch all rows from table
sql_query = 'SELECT * FROM bicycle_traffic'
bicycle_traffic = pd.read_sql_query(sql_query, connection)

# Remember to close the connection when you're done
connection.close()
```

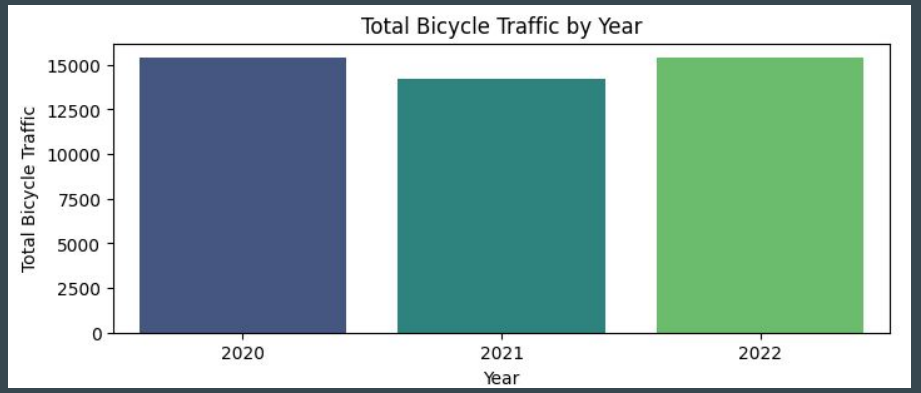
```
total_traffic_by_year = melted_data.groupby('Year')['Bicycle_Traffic'].sum().reset_index()

# Bar Chart: Total Bicycle Traffic by Year
plt.figure(figsize=(12, 6))
sns.barplot(x='Year', y='Bicycle_Traffic', data=total_traffic_by_year, palette='viridis')
plt.title('Total Bicycle Traffic by Year')
plt.xlabel('Year')
plt.ylabel('Total Bicycle Traffic')
plt.show()
```

```
# Define a custom color palette
custom_palette = sns.color_palette("muted", 13)

# Bar Chart: Bicycle Traffic Comparison for All Streets Over the Years with Custom Colors
plt.figure(figsize=(16, 8))
sns.barplot(x='Year', y='Bicycle_Traffic', hue='Street_Name', data=melted_data, ci=None, palette=custom_palette)
plt.title('Bicycle Traffic Comparison for All Streets Over the Years')
plt.xlabel('Year')
plt.ylabel('Bicycle Traffic')
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')  # Adjust legend position
plt.show()
```

Exploratory Data Analysis







Introduction

Data Sources

Data Engineering Pipeline

Exploratory Data Analysis

Results & Future Work

Results & Future Work for Project 1

Insights

1) May, June and July has the highest number of accidents in both the year.

2) Maximum accidents take place between noon 12 till evening 7.

3) The majority of accidents in Cologne caused slight injuries (7,956), followed by 1,337 incidents resulting in serious injuries, and 41 accidents leading to fatalities.

4) Collisions with turning or crossing vehicles are the most common accidents in Cologne, with 2,412 cases, followed by other accident types at 2,024 cases.

5) Daylight conditions witnessed the highest number of accidents in Cologne, totaling 6,989 cases, significantly surpassing accidents during dark (1,883 cases) and dusk (462 cases).

Future Work

Incorporating weather data, urban development information, or even socio-economic factors may provide a more holistic understanding. Moreover, expanding the dataset to include a more extended time frame or incorporating real-time data could capture dynamic changes in accident trends and contribute to more robust conclusions.

Results & Future Work for Project 2

Insights

1) In all the years the street with most bikers has been : Venloer Straße

2) And the street with the least bikers has been : Vorgebirgspark

3) The bicycle traffic has remained constant over these years

Future Work

In the future, this project aims to analyze recent data, uncover seasonal patterns, explore demographic influences, and correlate trends with infrastructure changes. Machine learning models and user surveys will enhance predictive insights and qualitative understanding. Integrating real-time data and enhancing public engagement ensure a comprehensive approach, while ongoing efforts in data quality improvement maintain accuracy and reliability.

Thank You