SAN: Final Assignment and Work Plan

Adrian Filcík, Adam Hrabovský, Zuzana Világiová, Matej Vojtek

December 2023

1 Assignment

1.1 Question

Statistical analysis of road accidents in Slovakia provides crucial insights into road safety in the country. How can the available data, both on an annual basis and broken down by individual regions, allow for a deeper understanding of patterns and trends associated with traffic accidents?

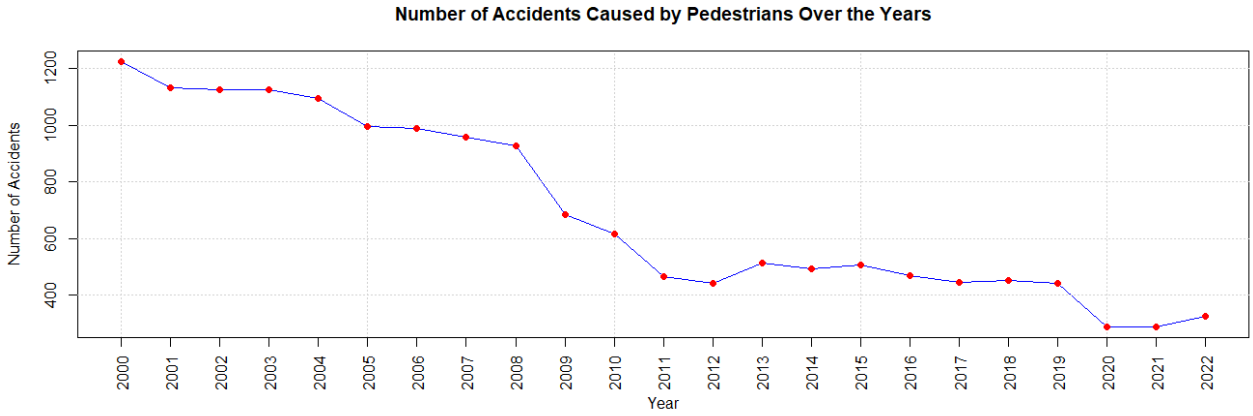

Figure 1: Example of Dataset Number of Accidents Caused by Pedestrians Over the Years

Research at the regional level adds value by helping identify regional disparities and specific characteristics. Each region has its own unique features that may influence accident rates, and understanding these local conditions is crucial for targeted interventions and strategies.

1.2 Data

For the analysis of accident rates, we will use publicly available data from the Statistical Office of the Slovak Republic:

DataCube

2 Work Plan

2.1 Specific Instrumental Questions

1. What does the data look like? Are all datasets worth including?
2. Is there any correlation between population density and the frequency of traffic accidents in different regions of Slovakia?
3. Are unsecured railway crossings associated with a higher rate of traffic accidents?
4. Is there anything that can be said about the average income in different regions and its connection to accident rates?
5. Is there any relationship between fuel prices and the frequency of traffic accidents within Slovakia?

6. How is the road infrastructure (including intersections and traffic lights) related to the frequency of traffic accidents?
7. Is there a relationship between the number of motor vehicles and the frequency of traffic accidents?
8. How does the death rate due to road traffic injuries change over time, and is there regional variability?
9. How is unemployment in different age groups related to the frequency of traffic accidents?

2.2 Datasets

For data analysis, we will use publicly available Accident datasets. The accident data is categorised into subsets, e.g., caused by pedestrians, incidents resulting in slight injuries, and fatalities. Furthermore, the accident data is segmented into overall values for Slovakia and individual regions. All data is collected from the year 2000 to 2022. The Statistical Office of the Slovak Republic gathered the available data using electronic data collection methods and household survey techniques. The specific datasets are following:

- 1.1.1. Demographic Balances of Population [om7001rr - om7006rr]
- 1.1.2.1 Total number of population [om7102rr]
- 1.1.3 Population Density [om7015rr]
- 1.2.3.1 Unemployed by age [pr3817qr, pr3117qr]
- 1.6.1.2 At-risk-of-poverty rate by regions of SR [ps3813rr]
- 1.6.1.5 Income per person - average income by regions [ps3807rr]
- 2.3.4. Average prices of fuels in the Slovak Republic [sp2039rs]
- 4.4 Accident frequency [do3806rr]
- 4.4 Length of local communications I. - IV. class in total (km) [do3805rr]
- 4.4 Road junctions controlled by traffic lights [do3805rr]
- 4.4 Total number of motor vehicles [do3803rr]
- 4.4 Survey on infrastructure situated on the roads of the 1st, 2nd and 3rd class - Secured railway crossings [do3804rr]
- 4.4 Survey on infrastructure situated on the roads of the 1st, 2nd and 3rd class - Unsecured railway crossings [do3804rr]
- 4.4 Railway transport - transport of goods and transport of passengers [do1005rs]
- 4.4 Railway transport - the indicators of performances and safety of operation [do1006rs]
- 4.4. Passengers transport [do2002rs]
- 7.1 Consumption of fuels, electricity, heat [en3001rr]
- 6.5.2 Death rate due to road traffic injuries [om3403rr]

2.3 Answering IQs

1. Data Gathering and Categorisation
    - Begin by assembling all the chosen datasets relevant to the study
    - Provide a detailed description of their categorical divisions to establish a comprehensive understanding of the data landscape
2. Variable Analysis and Visualisation
    - Plot all the variables under analysis against accident
    - Uncover initial trends and correlations

3. Statistical Testing
    - Apply a t-test or ANOVA test to assess the significance of differences between groups within the dataset
    - Conduct a correlation test to quantify the strength and direction of relationships between variables
4. Model Fitting
    - Establish a mathematical representation of the relationships identified in the data, contributing to a deeper understanding of underlying patterns

By following this systematic approach, we aim to extract meaningful insights from the datasets to evaluate statistical significance.

2.4 Risks and Limitations

- *How could confounding affect our results?*

- *Are there any significant irregularities in the dataset that should be addressed?*
  For example a large-scale event that introduces changes the data which cannot be explained by simple statistical methods.
- *Is there any data missing? How will this affect the results?*