Explantory Data analysis

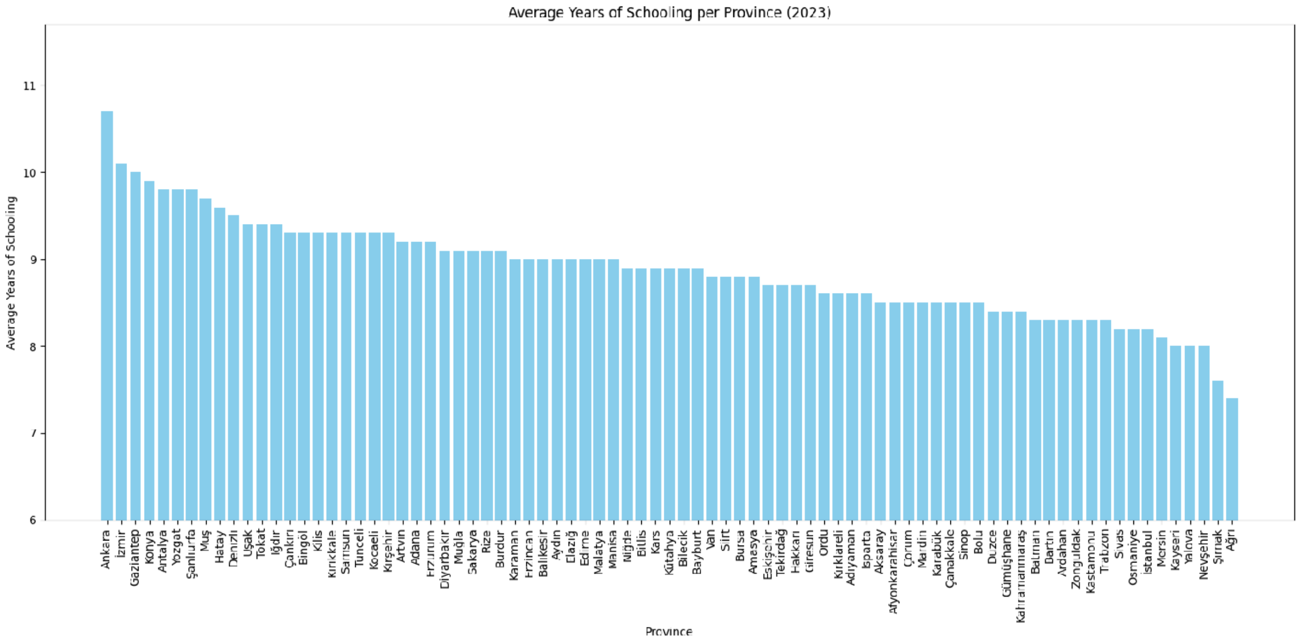
ibrahim suat gürsoy

Introduction

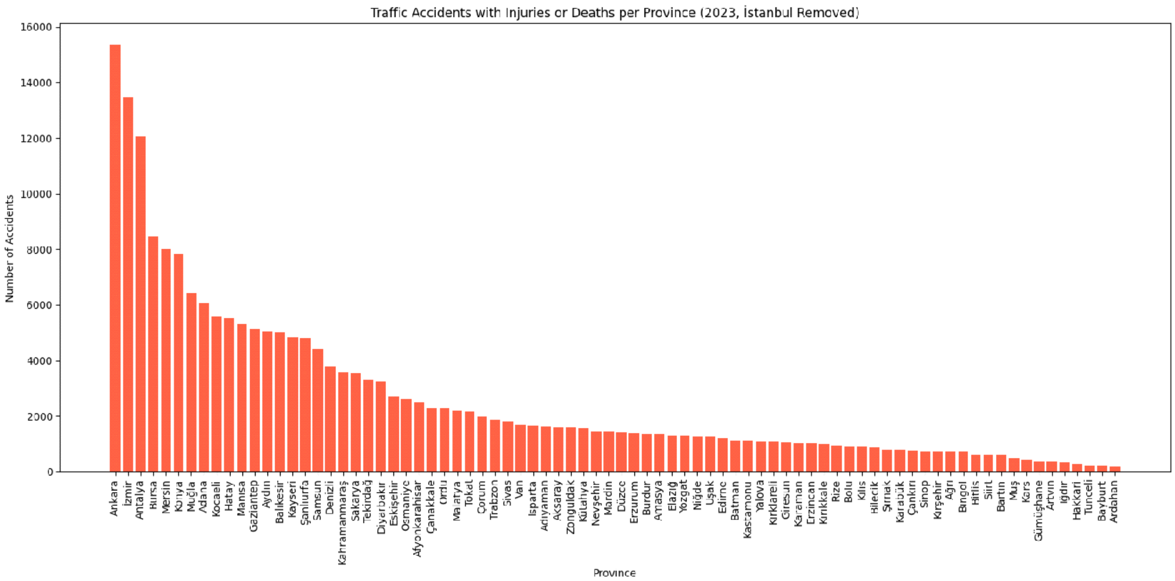
In this analysis I tried to correlate traffic accidents with the amount of time spent with schooling in order to find if there is connection/correlation between these stats. You can find the data I used for hypothesis testing(except population data it will be here later) and my data manipulation methods to understand if there is correlation.

DATA

Average education length in cities(check github repository if its not shown properly, Y-value is clamped to 4 in order to show differences better)



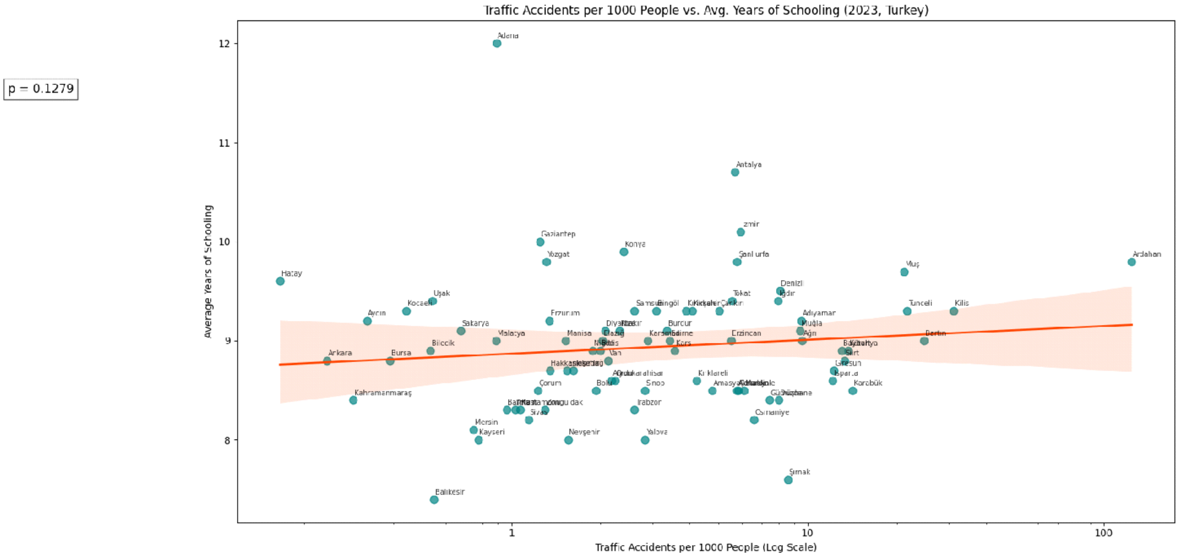
Traffic accidents per city(istanbul ommited).



Data Manipulations

Istanbul is removed from the dataset because of heavy immigration it gets. Accidents is also divided by the population in my analysis to eliminate population size based discrepancies. Data provided by TUİK doesn't follow any data representation models(like comma separated data) so I used regex to reformat the data(fixed it may 30). There were some outliers in the data that made plotting it difficult, i did not want to exclude that data so, i made the accidents per thousand axis logarithm based because when I plotted the graph it was very left sided(or call it right tailed). I will consider the amount of cars and household income later to try to get an idea about how developed the city is.

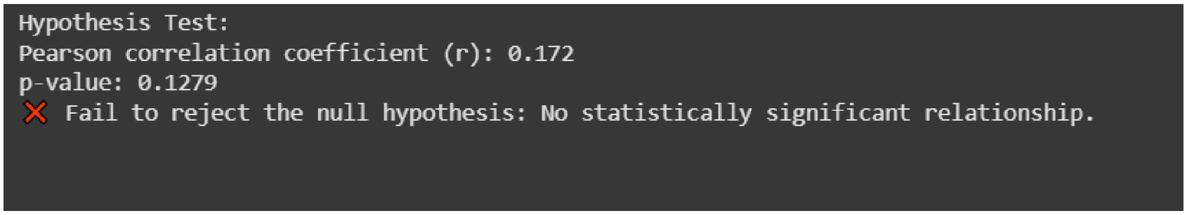
Plotting Results



We see a very distributed graph which suggests a lack of relationship between education levels. We will consider this distribution later in the hypothesis testing.

Hypothesis Testing

Null hypothesis for our case is, average Education time(i considered it same thing with education levels) does not effect traffic accidents in a given city. Alternative hypothesis is average education time does effect traffic accidents in a given city. In order to determine if we need to do hypothesis testing with significance = 0.05. If we use power of python to determine this results is:



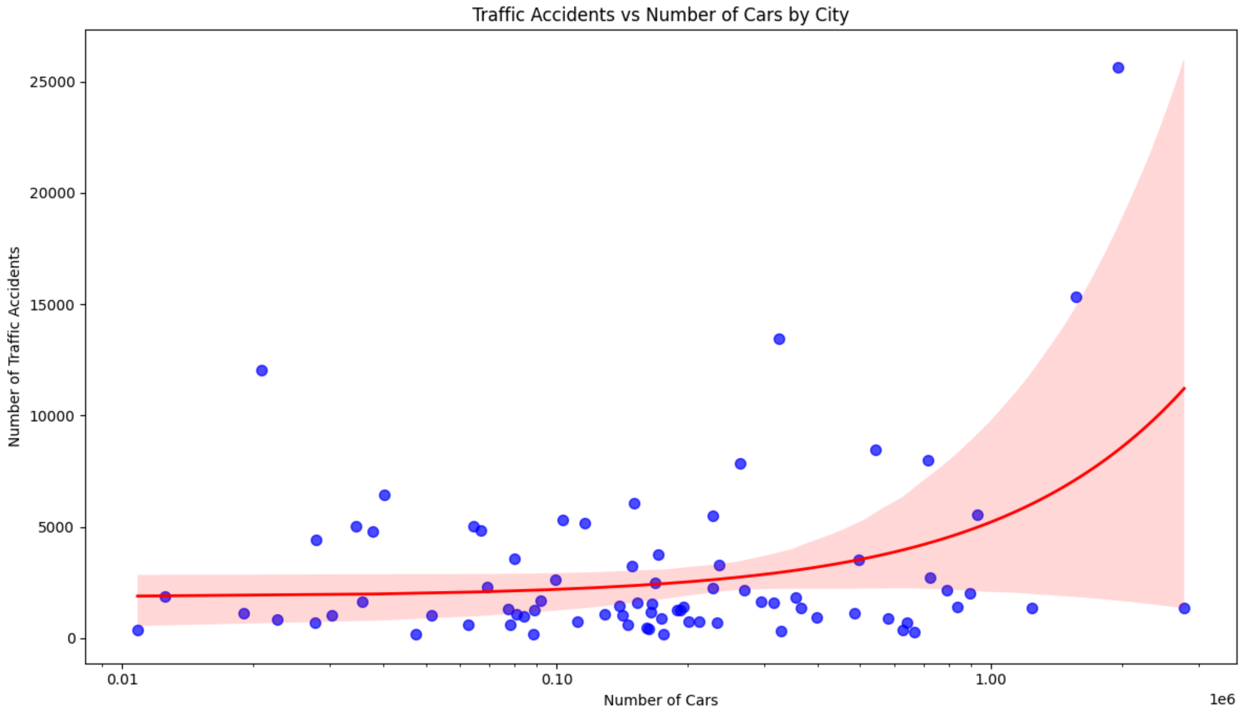
```
Hypothesis Test:
Pearson correlation coefficient (r): 0.172
p-value: 0.1279
❌ Fail to reject the null hypothesis: No statistically significant relationship.
```

Results

We can safely assume education time(levels) does not impact the amount of traffic accidents in Turkey. How much the data deviates from the trendline is concerning because it suggests some of the cities have low collision rates but other cities don't which means there are a lot of steps to take to reduce traffic accidents but it is being done in some cities not in every city. Later i will take other aspects of the city in the machine learning part where we use learning or regression algorithms to try to find an attribute that affects the accidents.

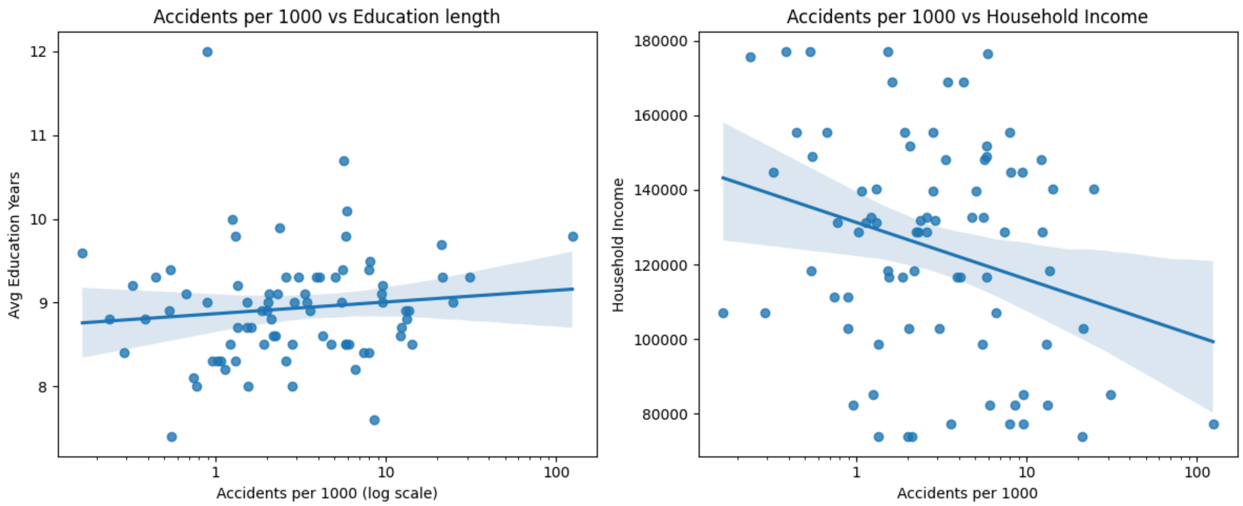
Machine Learning

I used many machine learning algorithms to understand what is affecting traffic accidents, from linear regression to XGBoost but the only outcome that made sense came from random forest regression. I will come later to why it was the only one that was giving somewhat of a prediction. Before we found out there was no significant correlation between education and traffic accidents, this fact hinders our ability to do machine learning on our data. TUIK provides information about most of the cities as an interactive map on their website, however i was not aware of the fact they didn't provide files for most of the data shown. I scrambled whatever I could from their portal to at least make sense why some cities had higher traffic accident rates. This data was the average household income and number of registered cars in cities.



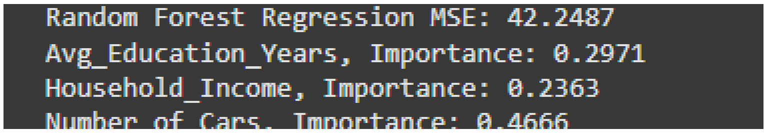
r = 0.38, p = 0.0002

There is a corelation between number of cars and accidents but it is not strong.

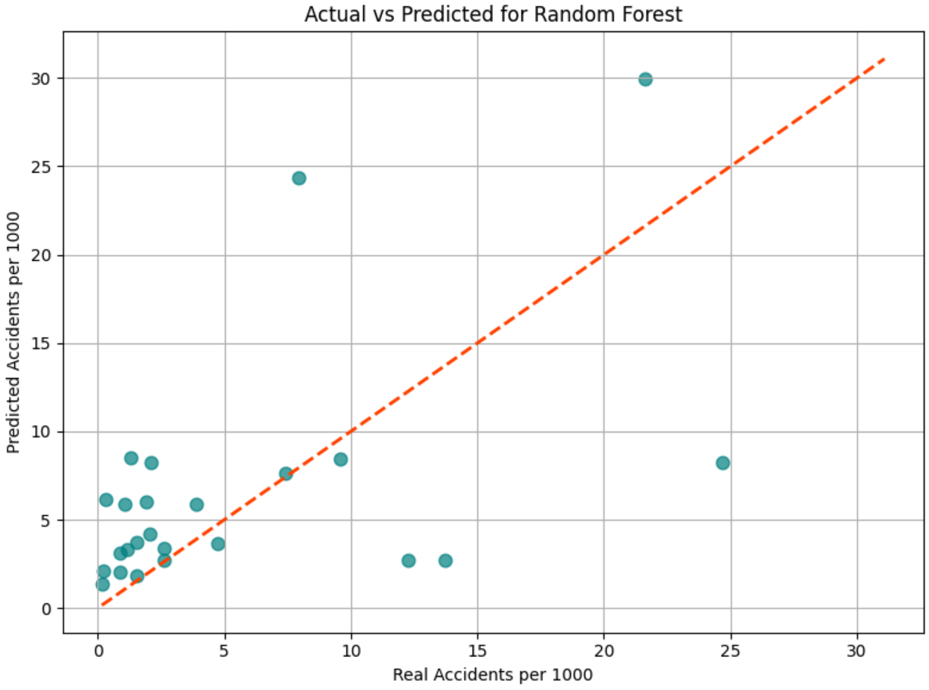


There is little to no correlation for these 2 factors.

My guess is for multilinear regression there is not enough data points as Turkey only have 81 cities. The result from any linear regression was over 150 MSE. which is high for our case. When random forest was used to train and predict the data MSE dropped to 40. It is still relatively high for a working “prediction” but it hints us that there is a non linear relationship between our 3 attributes, probably the number of cars is the biggest effect. Here is the graph for predictions for a %30 sample size.



```
Random Forest Regression MSE: 42.2487
Avg_Education_Years, Importance: 0.2971
Household_Income, Importance: 0.2363
Number of Cars, Importance: 0.4666
```



Conclusion and limitations

Real cause of traffic accidents is hard to point out using the data available. Our hypothesis was that education effects traffic accidents but we found out it was not the case for Turkey. Machine learning methods struggled to find a prediction for our cities because the level of education, household income or number of cars per 1000 people is not strongly correlated with traffic accidents. Biggest limitation for this project was data, TUIK provides abysmal files. Files are not annotated properly they don’t use the csv standard, there is data visually on the website but you can’t download it, even if you can download it it has missing data points all the time. Until a better dataset is found, the real cause of traffic accidents on a data level remains a mystery for us.