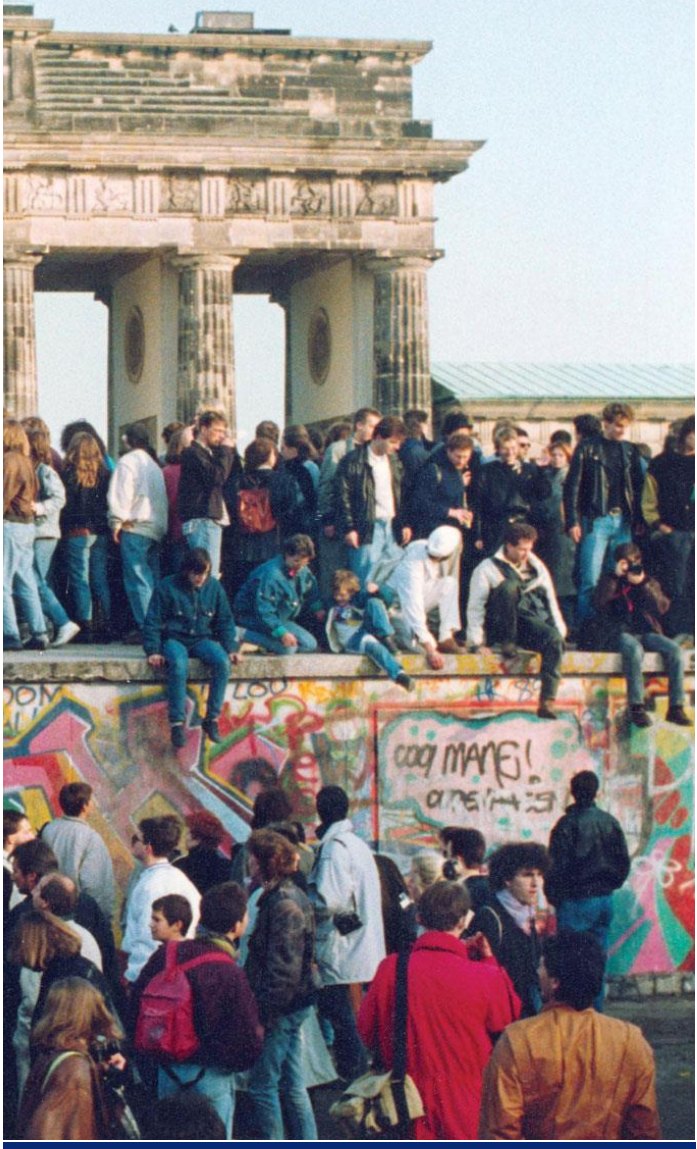
The East-West Divide

An investigation into the prevalence of economic disparities in Germany across historical lines post reunification.



MAY 7

315 Data Mining

Authored by: Harry Pines _ 11578059

1. Introduction

In 1990, the German Democratic Republic (GDR), formerly the most successful economy in the eastern soviet bloc had collapsed, with the fall of the berlin wall a year prior, the state merged with the Federal Republic of Germany (FRG) to form what is now the unified Germany we know today. At the time of the merger, there were significant economic disparities. Through socioeconomic merging under the West German mark, the divides have slowly faded across the country, the average lifespan and education of the east and west German is now at parity but the underlying question of how well the crack has truly healed is up for debate. In this report, I hope to examine that crack and visualize any traces it may have left.

What am I investigating?

First and foremost, this is an investigation into whether or not eBay listings can indicate the financial state of regions, If this investigation were successful then Id hope that answer would be yes.

Why bother?

I take great interest in history, and an opportunity to see if history can be still be seen through modern geographic data was an opportunity I could not turn down. I hoped to assess the state of German reunification almost 30 years on and see if there's still more work to be done bringing the two halves together.

What Challenges did I encounter?

With the obvious challenges that come with a dataset, much of the hard work went into learning how choropleth maps work and how they could be plotted, Doing so took me everywhere from python add-on packages to the European environmental agency.

- What did I achieve?

"Can eBay listings indicate the financial state of regions?"

While I did not necessarily find anything conclusive, there was a visible distinguishing divide between east and west Germany when looking at sale price and age. Whether this confirms or denies the hypothesis however will require further study and more data.

2. Data Mining Task

The task?

Here the task was to attempt to map the weighted averages to their corresponding regions to assess if there is a visually marked economic pattern that divides the map into similar regions to that of the former east and west German territories. with my sales dataset id hope to show that there would be a difference in the prices and ages of cars sold. I then mapped these differences to several choropleth diagrams.

- List all the data mining questions that you set out to investigate in this project.

- *what are the best metrics to use in an isolation forest to eliminate outliers?*
- *How can python libraries help me map data?*
- *Is there an underlying economic trend that distinguishes east and west Germany from one another, and if there is, are auto sales a good metric?*
- What are the common hurdles and mistakes that come with mapping averages?

Key Challenges

The data set, a lesson in counting chickens before they are hatched.

The largest problem with the data set is the type of listings it contained, the scrimper looked only at new listings which meant it was susceptible to all the eBay tricks employed by their users. Price data was no guarantee of success as joke listings --skewed my early attempts to extract results, there were prices from 10-million-euro ford Mondeos to a 2.1 billion-euro Toyota.
Obviously, these listings were never intended to sell and could safely be removed from the data set but there was difficulty in using the free or low initial bid listings.
These auctions typically start at nothing or some too good to be true value to attract bidders. While the other categorical data like mileage and age are most likely fine in these listings, their prices could not be used. Had the skimmer only used completed or sold listings, this might have been avoided.

Age of a car, from archeology to the distant future

Another one of the flaws, this time at the fault of eBay motors was the lack of a hard limit for registration input. It stands to reason that no car for sale could predate the invention of the car or have time traveled from the distant future to be listed here in 2016 but that became a serious issue when plotting the relative ages.

3. Technical Approach

Addressing the challenges

The Dataset, count something else instead

the solution to this was to install some hard cutoffs for the prices (no million-dollar cars or freebies) and then utilize an isolation forest to attempt to remove the other outliers. This succeeded in removing almost 100 thousand unusable price metrics.

Age of a car, reassessing fact.

I had cars dating from the birth of Christ to the dawn of the year ten thousand. Once again, these listings had other metrics that I did not want to lose from the data set and so a correctional plan was invoked.

Firstly, I isolated the reasonable time listings and calculated their weighted average and standard deviation. rounding up I found the weighted average car was typically registered in 2004 and varied by around 8 years. I then took the offending listings and populated their ages with a random age between 1996 and 2012 or within the standard deviation for the other cars.
doing so did not significantly alter my other categorical metrics at the end and the generally observed trend continued in the average age of cars across postcodes as is discussed in my results.

The Approach

First I converted my CSV into a data frame, I removed the unneeded categories, capped the price at 10 million and created a new column for the rounded postcode, I then find the weighted average and standard deviation of the filtered years and map the bad ones with new values. I then run my data frame through the isolation forest from skLearn with contamination of 12%. Then I can compute and map the averages or weighted average to a choropleth map.
The weighted averages required additional work. First, a count was needed from each code and then tied to the associated listing, they then needed to be grouped by their rounded postcode value and then computed.

```
def computeWeighted(dataset, KEY,msg):
    averagePrice = np.zeros((len(dataset),),dtype=object)
    for i in unique_postcodes:
        newDf = pd.DataFrame(dataset[i])
        mode = 0
        try:
            mode = newDf.avgCode.mode()[0]
            newDict = {'mean': newDf[KEY].mean(),'weight' :len(newDf.index), 'code': mode}
            averagePrice[i] =  np.array(list(newDict.values())).astype(float)
        except IndexError:
            averagePrice[i] = 0
```

Snippet of weighted average filtering code:

In order to determine which isolation forest metric worked best, I tested the following:
-Price vs Mileage
-Price vs Age
-Mileage vs Age

-Manually filtering age.

Price and mileage resulted in the best data so that has been the main outlier filtering methodology employed. The resulting average/postcode data frame was then merged with a shapefile of German counties. This one was found through the internet archive and is dated circa 1999, there have been some complaints that it is now inaccurate but unfortunately, I was unable to get the geojson files from the European environmental agency working and thus this has been the best I could come up within the timeframe.

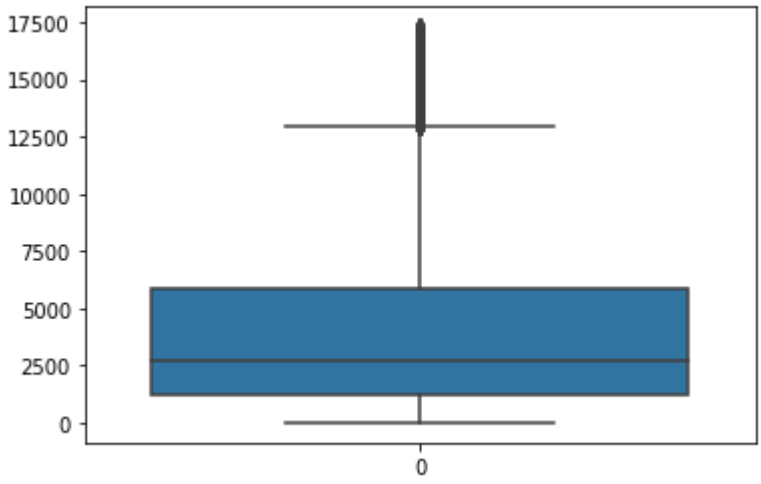


Figure 1, Box plot of Price-Mileage Price distribution

4. Evaluation Methodology

The Metrics.

The primary metric of measurement has been choropleth maps. these were primarily chosen as it allows readers to draw their own pattern from the data without having to tell them explicitly, working with geographic data meant plotting it visually would be essential particularly in support of my hypothesis.
the data classification going in involved a combination of weighted and unweighted means. Individual postcodes were combined with their hundred nearest neighbors to extract a better picture from the data.

Answering the questions

- *what are the best metrics to use in an isolation forest to eliminate outliers?*

Price was a critical metric as this had the widest range of bad results after correcting the year data.
Combining it with mileage became the best metric to use in the isolation forest.

- *How can python libraries help me map data?*

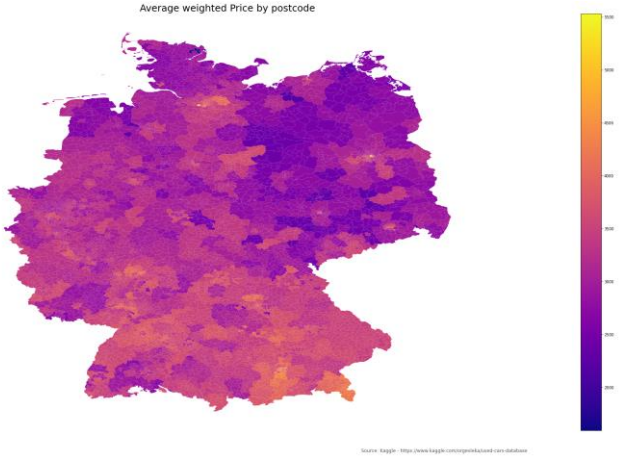
MATLAB is an immensely helpful library but if I had more success with finding a geojson, I would have made better maps with PLOTLY and GEOPANDAS. Getting the map might be the hardest part.

- *Is there an underlying economic trend that distinguishes east and west Germany from one another, and if there is, are auto sales a good metric?*

There might be, but there are many other factors at play to be certain. Other sources support that there is some economic disparity even now between east and west Germany but whether car sales can show it is up for debate.

- What are the common hurdles and mistakes that come with mapping averages?

Knowing when somethings wrong with your map is essential to good mapping. Early versions of the price data didn't take into account the wilder upper figures and left maps almost devoid of color. As is detailed on the next page, weighing the data is also important to ensure you are balancing it across the board. Where unrelated influences might skew or hide the pattern we are looking for.

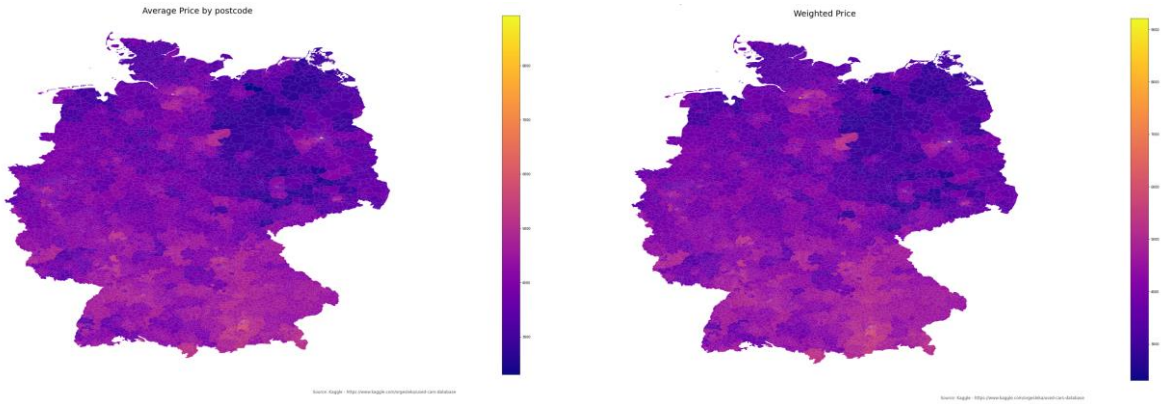


Not excluding € 0 price tags from the isolation tree left it including most free listings and leaded to more wildly varying data.

Uncapped isolation tree measurement:

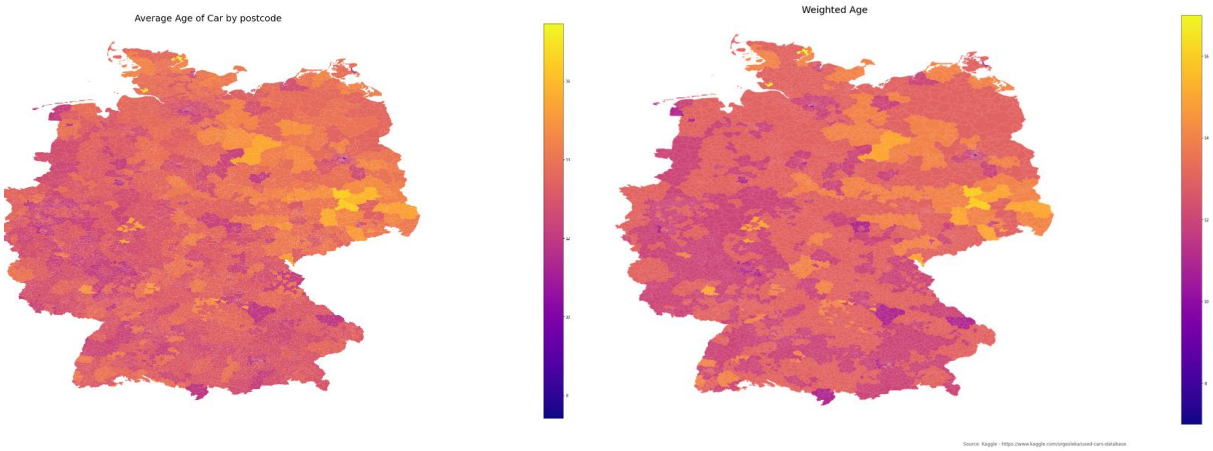
Weighting the data,

To ensure the accuracy of my pricing data I went about trying to compute the weighted average based on each postcode. Surprisingly, I did not find any noticeably visible change in price data when compared to its weighted equivalent,



Fig(2) Average price data in € (left), Weighted Average price data in € (right)

However, A marked difference was found when the same was done for the average ages of cars sold.



Fig(3) Average age data in years (left), Weighted Average age data in years (right)

The Dataset,

The data set consisted of approximately 370 thousand used car listings from German eBay in the month of March 2016, it was scrimped and submitted to Kaggle where it became a free to use source. while certainly large in scope and depth, there were challenges involved when working with optional user headings.
While its challenges have already been discussed in detail, to reiterate the greatest problem has been the reliability of the users to place correct data in their listings.

While I acquired the data through Kaggle, they collected it through skimming eBay motors.

Results,

After sorting and cleaning the data, calculating the respected deviations, weighted and unweighted metrics and mapping to their respective choropleth maps, here are the 3 most supporting of my hypotheses,

A rough outline of the FGR and GDR has been overlaid and major cities have been marked for the convenience of the reader,

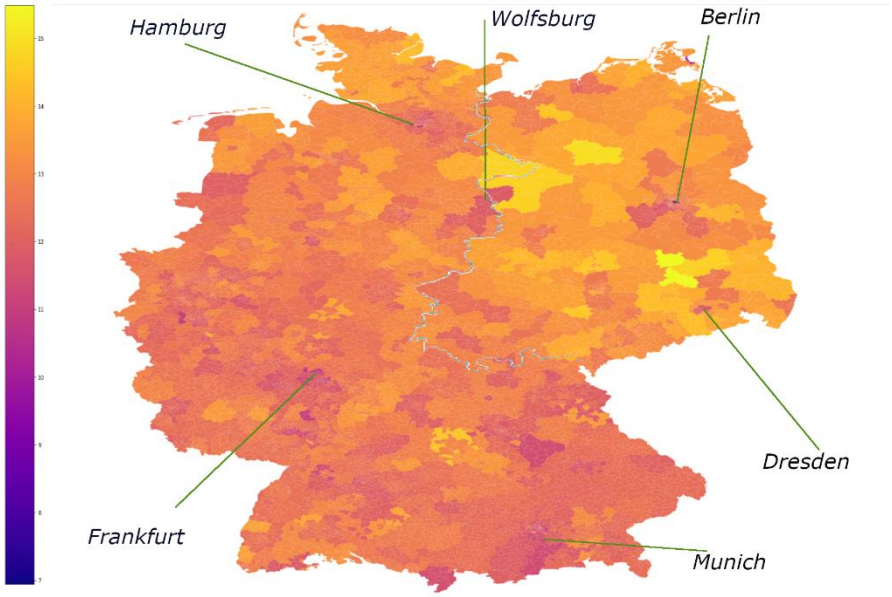


Fig 4) Average Age Data in years

Age

"On average, the cars sold in the area bounded by the former GDR are older than cars sold in the former FGR, East Germans are holding onto their cars longer than their western counterparts."

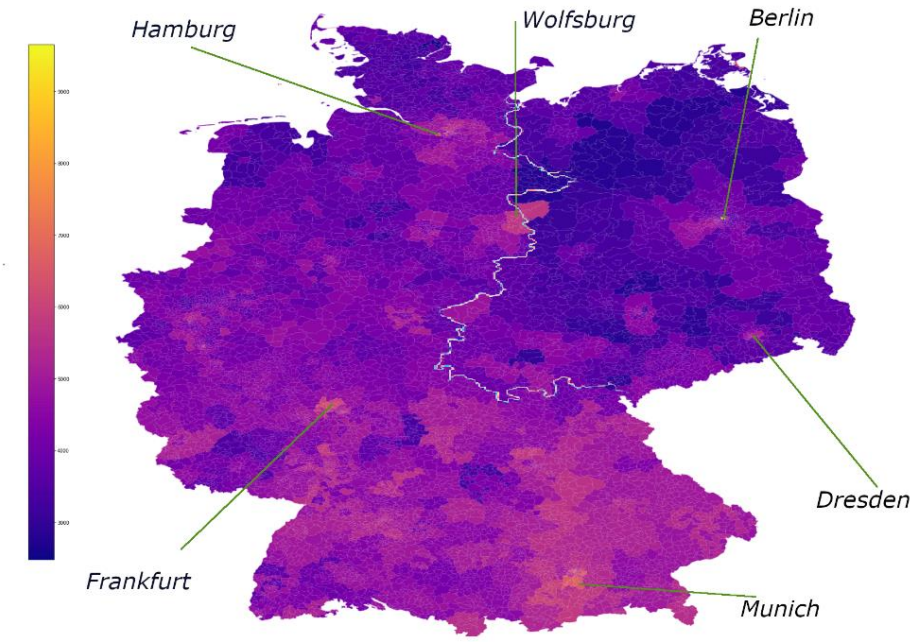


Fig 5) Average Price Data in **€**

Price

"The cars listed in east Germany are priced less on average than cars sold in the west. This would suggest that they aren't buying as many luxury cars as the west"

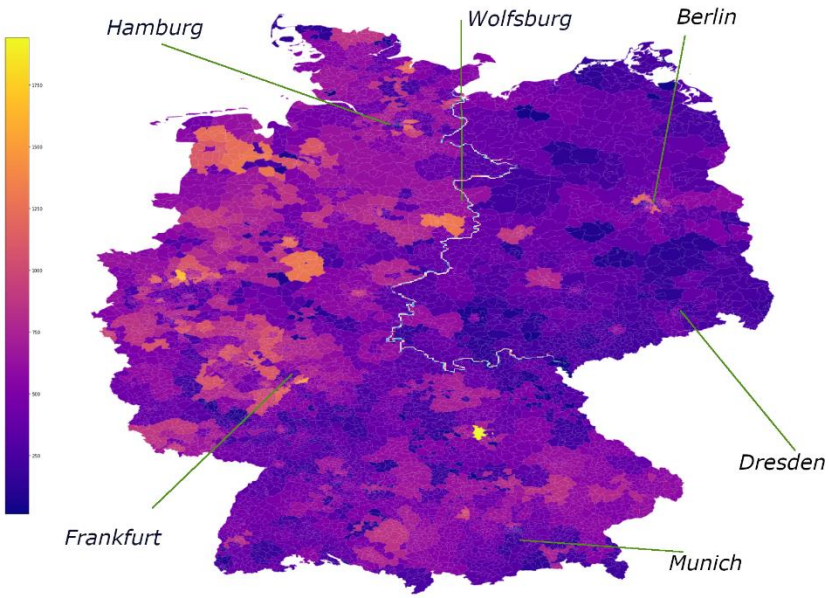


Fig 6) Average Listing Data

Listings

There are less cars listed in east Germany, considering the userbase is evenly distributed, it would indicate that east Germans have fewer cars to sell or sell them less frequently."

What did I discover?

First and foremost, the results suggest that there is indeed a difference in the general age and sale price of cars that closely follow the original borders of un-unified Germany. The average selling price was lower in the east than the west and the average age of cars sold was greater in the east than the west. There were also comparatively fewer listings in the east than the west.

The best conclusion that can be made from the scope of this report is that rural areas are typically buying cheaper older cars on the used market than their urban counterparts and this trend is more evident in formerly GDR territory than FRG. This could support the conclusion that the former GDR region is at a financial disparity compared to the FRG, but more time and data would be needed.

What worked and why?

The price, listing, and age data most clearly backed my hypothesis and supported earlier assumptions that were already made about the financial state of Germany. The most expensive newer cars were concentrated in the wealthiest parts of Germany suggesting that car data might be a good indicator of the prospective wealth of a region.

What did not work and why not?

I had great difficulty removing outliers from my data set, depending on the isolation method used I ended up with wildly different maps, they all supported my hypothesis to a degree, but the variance did not help my confidence.

the methods used to display my data also varied wildly, I had initially set out to learn the PLOTLY libraries but that ultimately fell through due to my inability to acquire useful postcode maps of Germany.

6. Lessons Learned

What I learnt?

In this project, I gained a better understanding of the Pandas library and database manipulation. I also gained valuable experience in map-making through MATLAB and

even though I'd failed to use more modern methods like ".geojson" and PLOTLY, I was still able to create and utilize choropleth mapping in my results. Applying modern data to extrapolate arguments with historical context has made for a more interesting project and it is something I would love to continue researching.

Next time.

If there was a next time, I would have liked to employ more methods to normalize my data to give me greater assurances that the observed pattern is not fictional.
I would collect more recent eBay data but make sure I only look at sold listings to reduce my outliers and get a more consistent picture. I would also employ a data set of cars to lookup static data to fill missing fields like age or horsepower to get around the limitations of user-entered categories.

In conclusion, while I cannot prove my hypothesis with certainty, its one I hope to evaluate properly in the future, there seems to be a trend, at least visually and quantifying that trend is the next step.

Data-driven discussions are the forefront of the modern world but sometimes stepping back and working out how best to visualize data is just as important as the data itself. Even if I did not conclude anything significant, I learned how best to show it.

7. Acknowledgements

Stack exchange and Arnulf Christl for the shapefile,
https://gis.stackexchange.com/questions/25060/postal-areas-for-germany
Stack overflow for all my programming questions,