SUSAN WANG

Data Analytics Portfolio

PROJECTS

PIANO PRODUCTION	*Excel & Python*
SPOTIFY ANALYSIS	*Python*
BERLIN S-BAHN	*SQL & Excel*
CLIMATEWINS	*Machine Learning*

Pianos in the Digital Age

Market Analysis of Germany’s Piano Industry

Germany's Pianos

Context:

Germany is renowned for its production of premium acoustic pianos (upright and grands). The piano manufacturing industry is supported by the nation's strong musical heritage and music education system.

The digital keyboard entered the market in the 1980's. This study looks at Germany's piano market in the last decade to investigate how the rise of the digital keyboard has affected the production of acoustic pianos.

Key Question:

How has the digital keyboard and international trade affected the piano manufacturing industry?

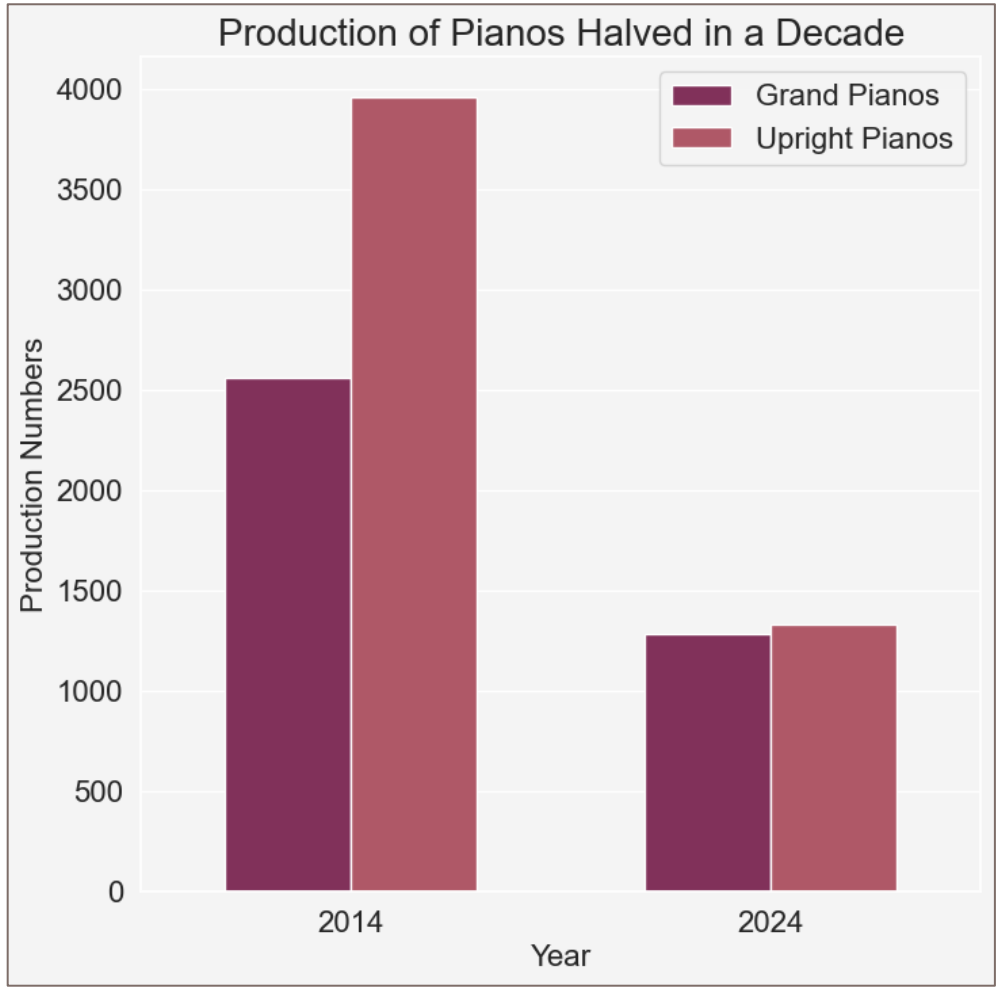
Data:	World Integrated Trade Solution (WITS) Destatis: GENESIS-Online
Skills:	Data cleaning and wrangling Exploratory analysis, Correlation testing
Tools:	Excel, Python

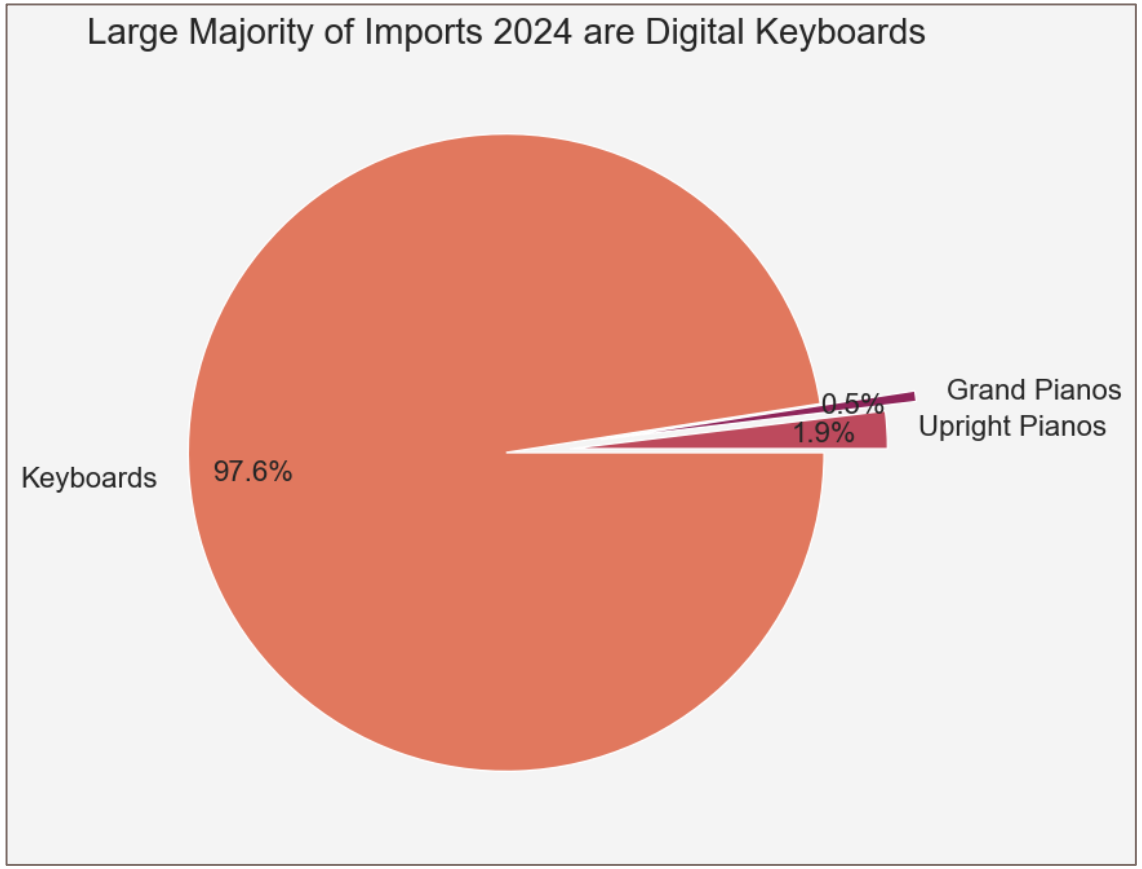
Process:

- Cleaning and wrangling data
- Exploratory analysis – finding trends
- Testing a hypothesis – correlations
- Summarize insights – infographic

Current Piano Market

The piano market includes import/ export of instruments as well as domestic production.





Significant observations:

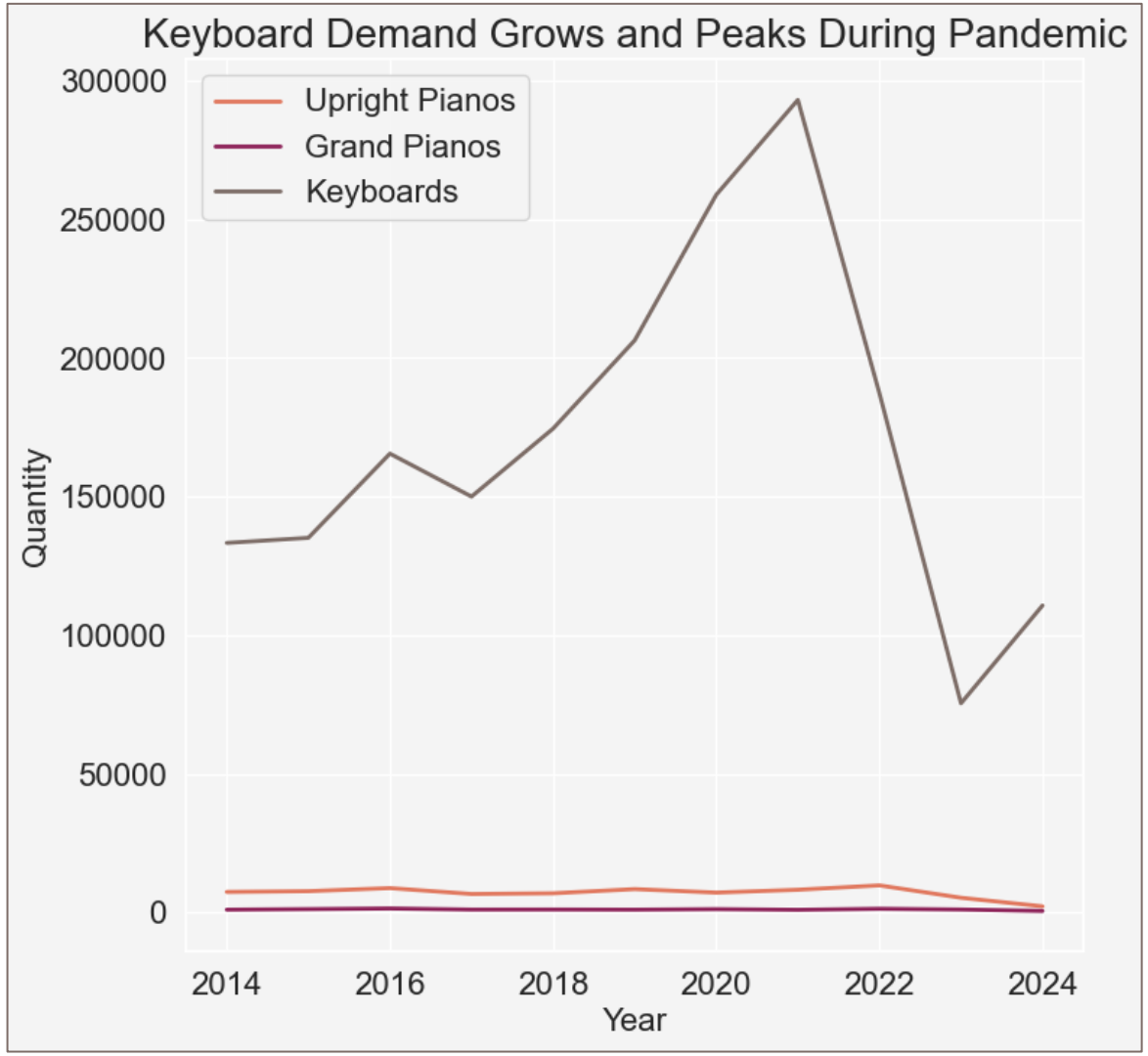
- The production number of acoustic pianos in 2024 is less than **half** of that in 2014.
- Digital keyboards account for **97.6%** of imported instruments.

Domestic Market

Domestic demand, calculated as **net import + production**, shows that digital keyboards dominate the market, peaking during the **pandemic** and falling dramatically after 2021.

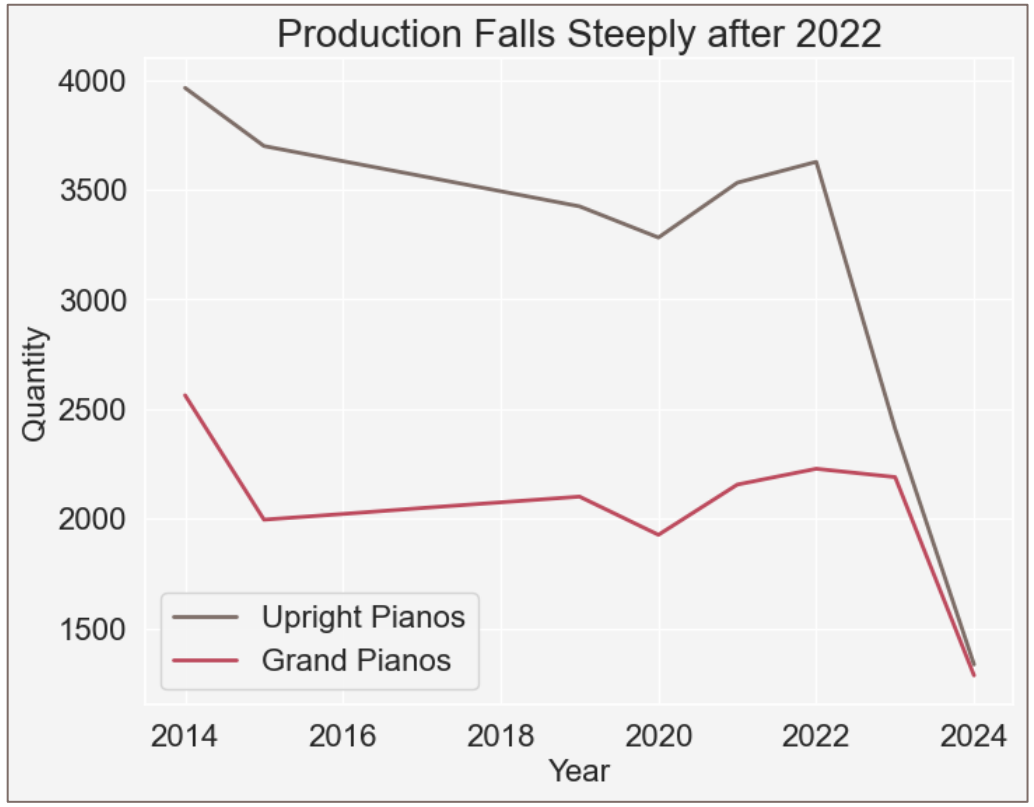
At **1/6** the value of the comparable upright piano, the digital keyboard is the most **affordable** option for consumers.

Germany exports **1.4x** more grand pianos than it imports, indicating strength in manufacturing and export market.

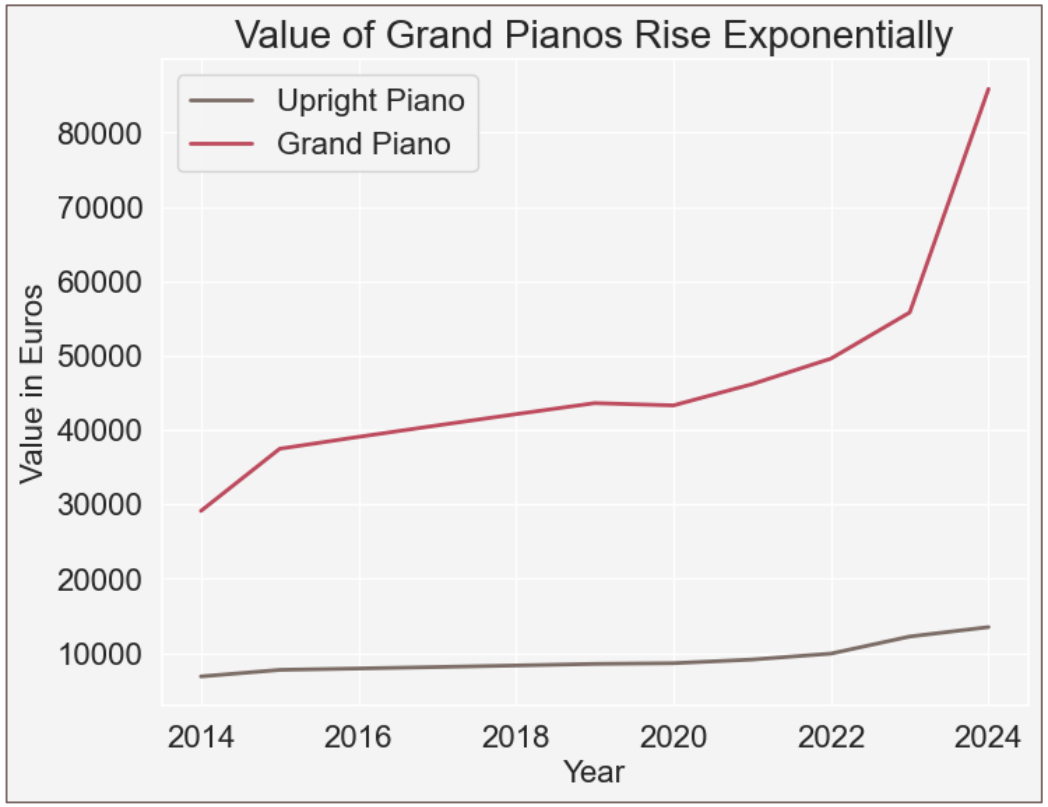


Quality over Quantity

In the same period as demand for digital keyboards was rising (2014-2021), production of acoustic pianos was already in decline. Yet, the most substantial fall in production happens after 2022. How did piano manufacturers stay competitive in the market?



Hypothesis: As production decreases, the value (or price) of pianos increases.



Correlation Coefficient

Grand pianos: **-0.81** Upright pianos: **-0.93**

A correlation test shows a strong negative correlation between production quantity and value, proving that manufacturers increase prices as production decreases.

Impacts and Consequences

Digital keyboards dominates the market as the most affordable option for consumers.

Quality of music education may suffer; cheaper digital keyboards offer broader accessibility to basic music education but limits capabilities for producing **authentic touch** and **refined piano playing**.

Piano manufacturers turn to quality over quantity: acoustic pianos become high-end premium products.

As piano manufacturers focus on the high-end market, they will trade in production numbers for luxury handcrafted instruments.

After 2022, there is sharp downward trend in both imports and production, indicating decreased demand.

With the decline of imports and production, there may be an opportunity for sales of **used pianos** to fill in the market gap for affordable acoustic pianos.

Conclusion

Project Assessment:

Goal:

- To explore the impact of digital keyboards on the structure of the piano market in Germany.

Challenges:

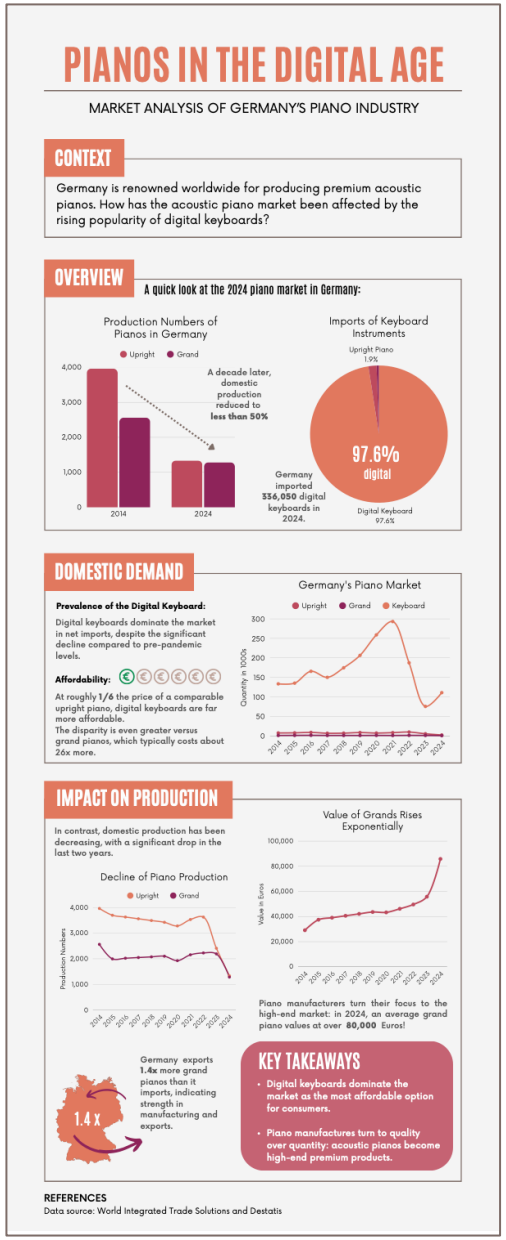
- Limited access to data, resolved through use of international trade and federal production numbers as proxy.
- Wrangling of variables to produce useful metrics for analysis, such as net import, export ratio and instrument value.

Results:

- Observed the growth and dominance of digital keyboards.
- Examined the impact of lower production numbers on increased value for acoustic pianos.

Reflection:

- In producing the results as an infographic, this project has taught me how to organize and present my findings in a concise and effective way.



Infographic created on Canva.com

View Project on GitHub

Spotify Music Analysis

Exploratory and Predictive Analytics

Overview

Context:

Spotify is one of the most popular music streaming apps today. There is ample data collected on artists and songs on its platform. An exploration of the top hits over two decades aims to discover trends and patterns of the most popular songs.

Key Questions:

- Where do most popular songs come from?
- Which audio features define a top hit?
- How has the music changed over the years?
- Can we predict how popular music sounds in the future?

Data:	Spotify Top Hit Playlist 2000 – 2023, Music Artists Popularity Data Set
Skills:	Data cleaning and wrangling, Exploratory analysis, Machine learning models, Dashboard creation
Tools:	Python, Tableau

Process:

- Preparing the data – cleaning and wrangling
- Exploratory visual analysis – finding correlations
- Regression analysis – testing a hypothesis
- Time series analysis – testing for stationarity
- Geospatial analysis – visual insights through mapping

Sourcing and Preparing Data

Primary Data Set:

- Sourced from Kaggle
- Top 100 hits per year on Spotify from 2000 – 2023
- 23 variables, including audio features such as *danceability, energy, key, mode, loudness, duration, tempo, valence, acousticness and danceability*
- Collected through Spotify API

Secondary Data Set:

- Sourced from Kaggle
- Data on more than 1.4 million artists
- Variables on artists, including *name, country, tags, and popularity*
- Collected from the MusicBrainz database and webscraping last.fm

Merged Data Set:

1. Data checked for missing values and duplicates
2. Wrangling procedure to prepare for merge
3. Merge on Artists' Name
4. Missing values in 'country' variable filled using assistance from ChatGPT

Exploratory Visual Analysis

1

Relevant variables were placed in a **correlation matrix heatmap** with the following results:

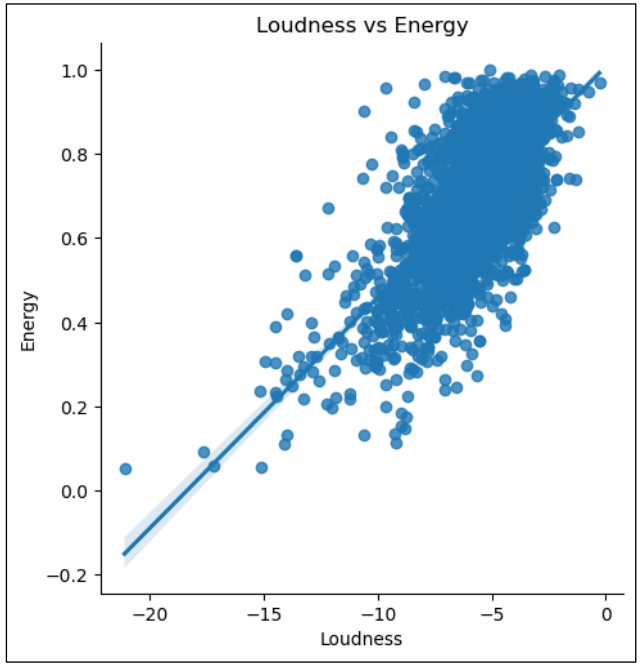
- Strong positive correlation (0.69) between energy and loudness
- Strong negative correlation (-0.55) between acousticness and energy

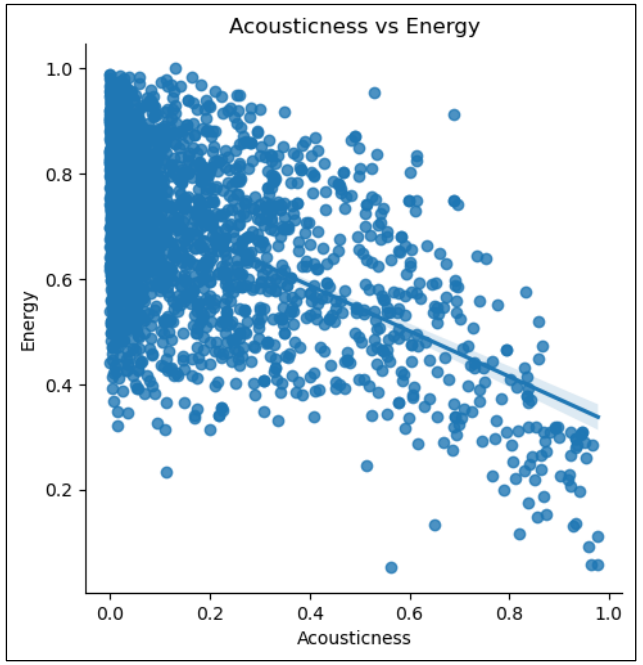
2

The correlations were visualized through scatterplots. The relationship between energy and acousticness was chosen for further exploration.

3

Hypothesis:
The more acoustic a song is, the less energy (or calmer) it feels.

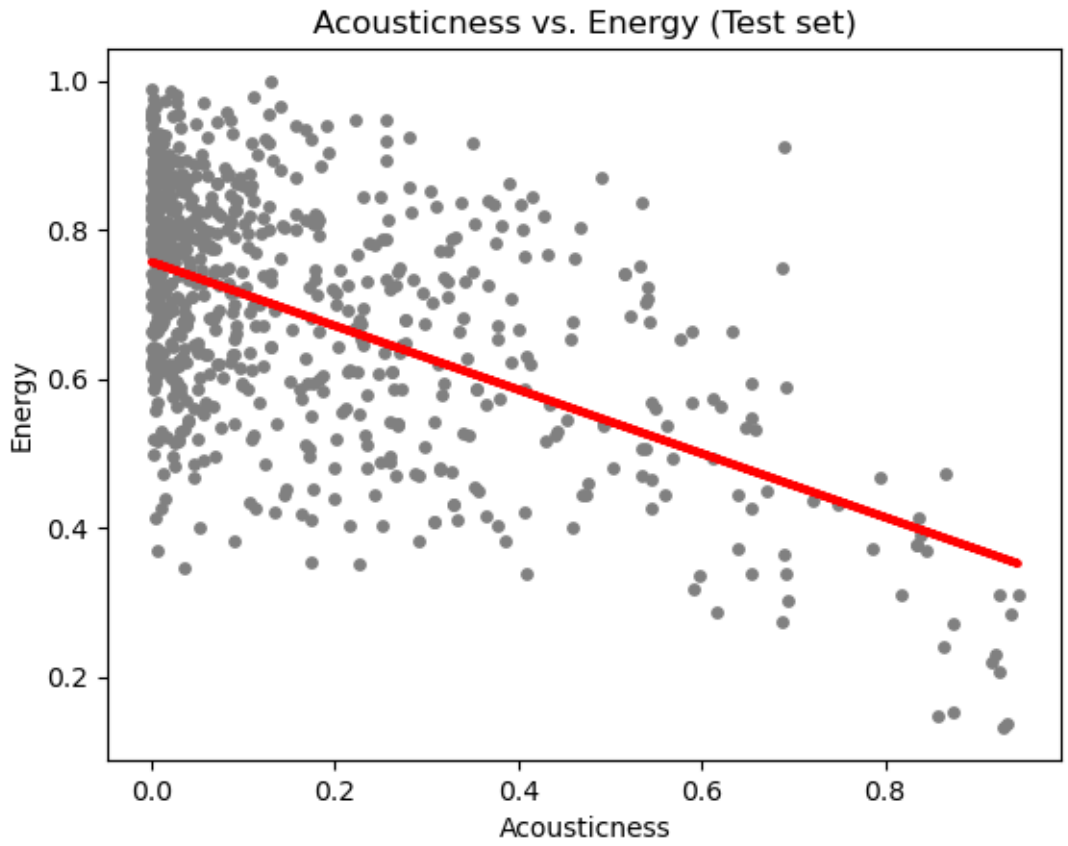




Linear Regression

The data was split into a training set and test set (70/30).
A **linear regression model** was applied to the training set based on the following hypothesis:

The more acoustic a song is, the less energy (or calmer) it feels.



The chart displays the results of the model on the test data set. The **model performance statistics** are as follows:

Slope: -0.42
MSE: -0.019
R2 Score: 0.309

- The negative slope confirms the negative correlation between acousticness and energy.
- The low MSE shows that the model's predictions are close to the actual values.
- The low R2 score indicates that this model **cannot** explain almost 70% of the variances in the data: the model is a **poor fit**.

Interpretation of results:

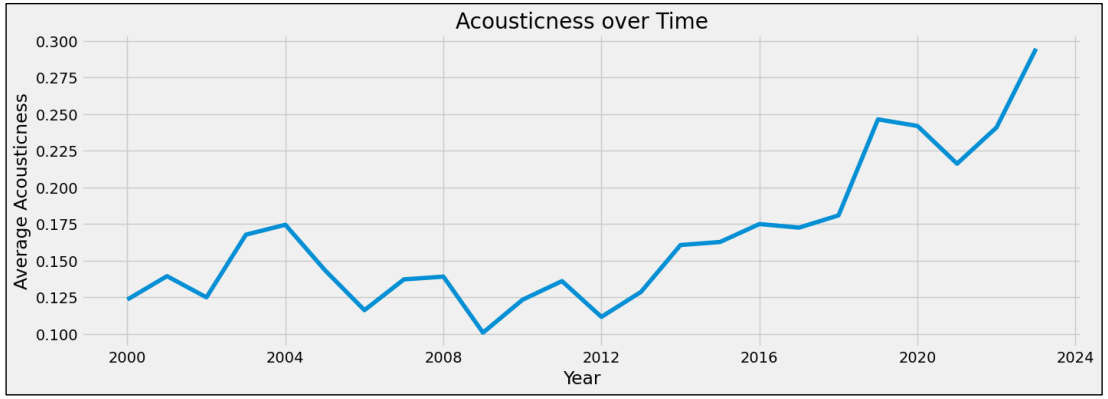
Acousticness alone does not explain energy level of a song. For the complexity of this case, a more advanced multiple linear regression model may prove a better fit.

Time Series Analysis

Acousticness over time:

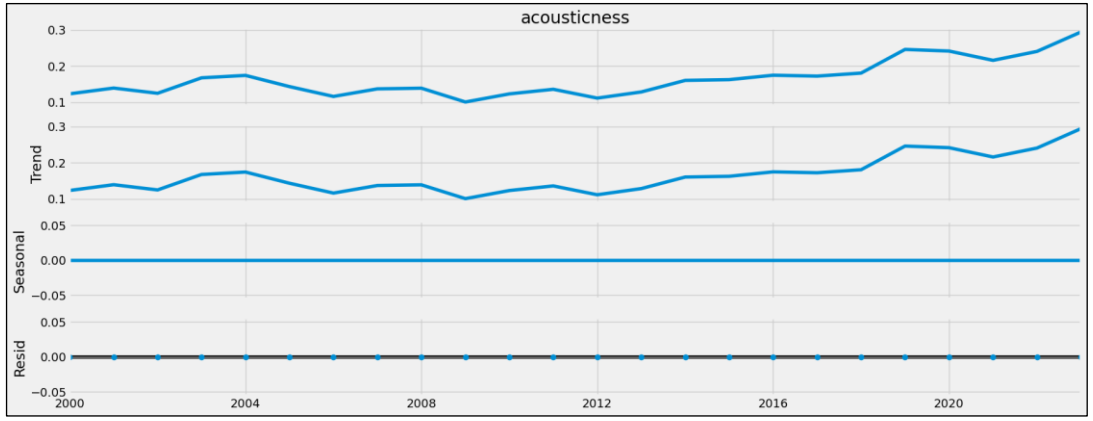
The introduction of electronics to music came only in about the last half century and is a relatively young genre in the long history of music.

Has the use of acoustic instrumentation continued to decrease since the beginning of this millenium?

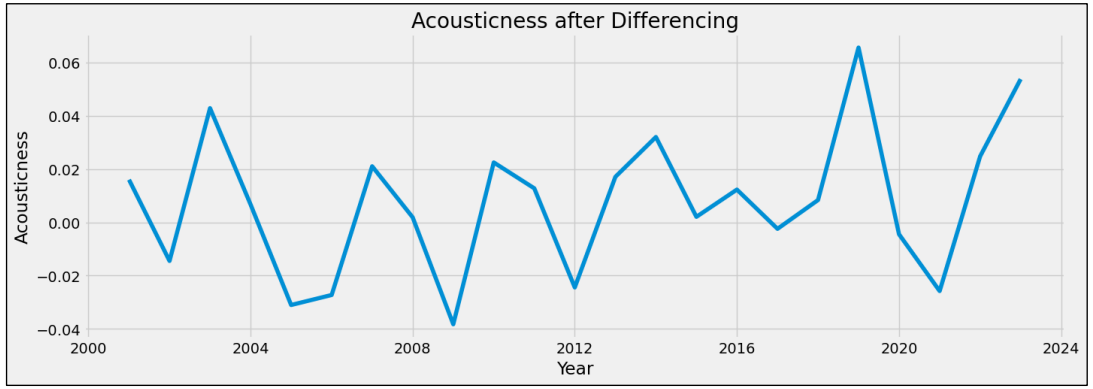


Surprisingly, average acousticness of the top hits since 2000 show a **rise in the last decade**. It seems that the electronic age may have peaked in the early 2000's and acoustic instrumentation is now making a comeback.

Stationarizing the time series:



The decomposition of the time series show a slight upwards trend, with no seasonality nor residuals.



Even after differencing, this time series could not be stationarized.

Conclusion: This time series is not suitable for forecasting.

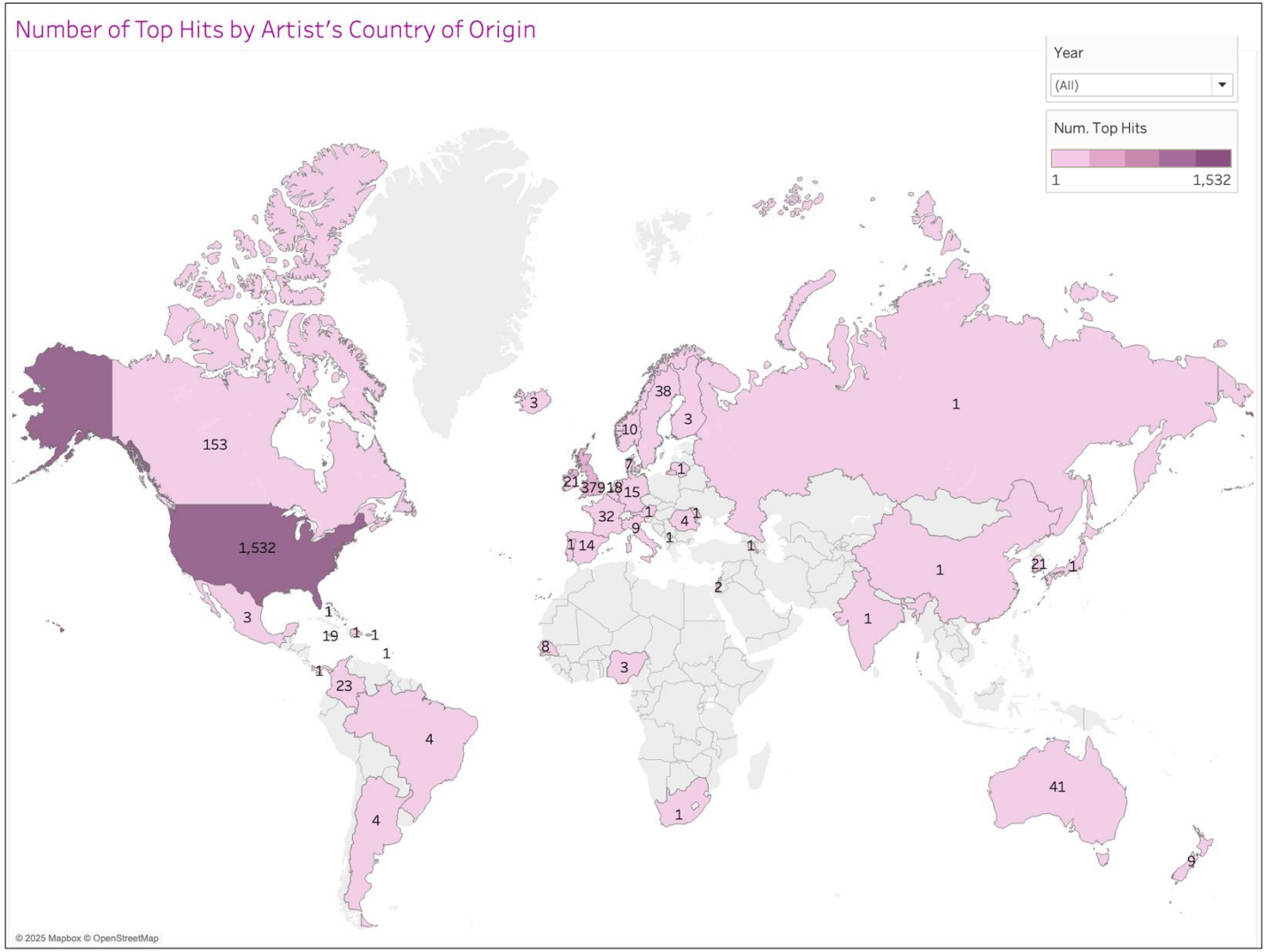
Geospatial Analysis

Where do the top hits come from?

Key takeaways from the geospatial analysis:

- The **USA** is home to the highest number of top hit artists.
- **English-speaking countries** (USA, UK, Canada and Australia) dominate in total number of hits over the last two decades.
- **South Korea** starts to gain significant numbers of top hits after 2019 and currently produces the highest number of popular artists in Asia.

Please click here to visit the interactive version of this map on Tableau.



Conclusions

Key Insights:

- American artists continue to top the charts in popular music.
- The last two decades have been a flourishing time for electronic (non-acoustic) music that are high in energy.
- Acoustic instrumentation has gradually returned over the last few years.
- The data doesn’t allow for forecasting, therefore we cannot predict how popular music will evolve in the future.

Project assessment:

The accuracy and reliability of the data on audio features are highly dependent on the specifications of Spotify's algorithms.
Ultimately, the data proved unsuitable for various models, such as linear regression, cluster analysis and time series forecasting.
Although these limitations created challenges in producing the insights to popular music I was hoping for, it reminded me that music is an art, and numbers cannot always capture the essence of art.

Next steps for further analysis:

- Perform an advanced multiple linear regression model on multiple audio features.
- Observe and compare audio features of top hits by various artist's nationalities.

The Tableau Storyboard and GitHub Repository are available to view:

Berlin S-Bahn Analysis Project

Transit Delays on Berlin’s S-Bahn in 2024

Overview

Context:

Berlin's S-Bahn is the rapid transit network connecting the city center with surrounding neighborhoods. With 6 lines and 10 starting stations, the S-Bahn recorded over 131,000 trips in 2024. This project analyzes those trips to identify the factors driving delays and cancellations throughout the year.

Objectives:

- Identify which S-Bahn lines and time periods experience the most delays
- Examine delay patterns by time of day, weekday, and month
- Investigate the impact of incidents, weather, and strikes on punctuality

Data:	Berlin S-Bahn Multi-Table Dataset
Skills:	Data cleaning and summarizing, database querying (CTEs, multi-table joins, aggregate functions), reporting
Tools:	PostgreSQL, Excel Reporting, PowerPoint

Process:

- Understanding the data structure
- Exploring and preparing the data
- Answer business questions through queries
- Report findings and deliver insights

Exploring the Data

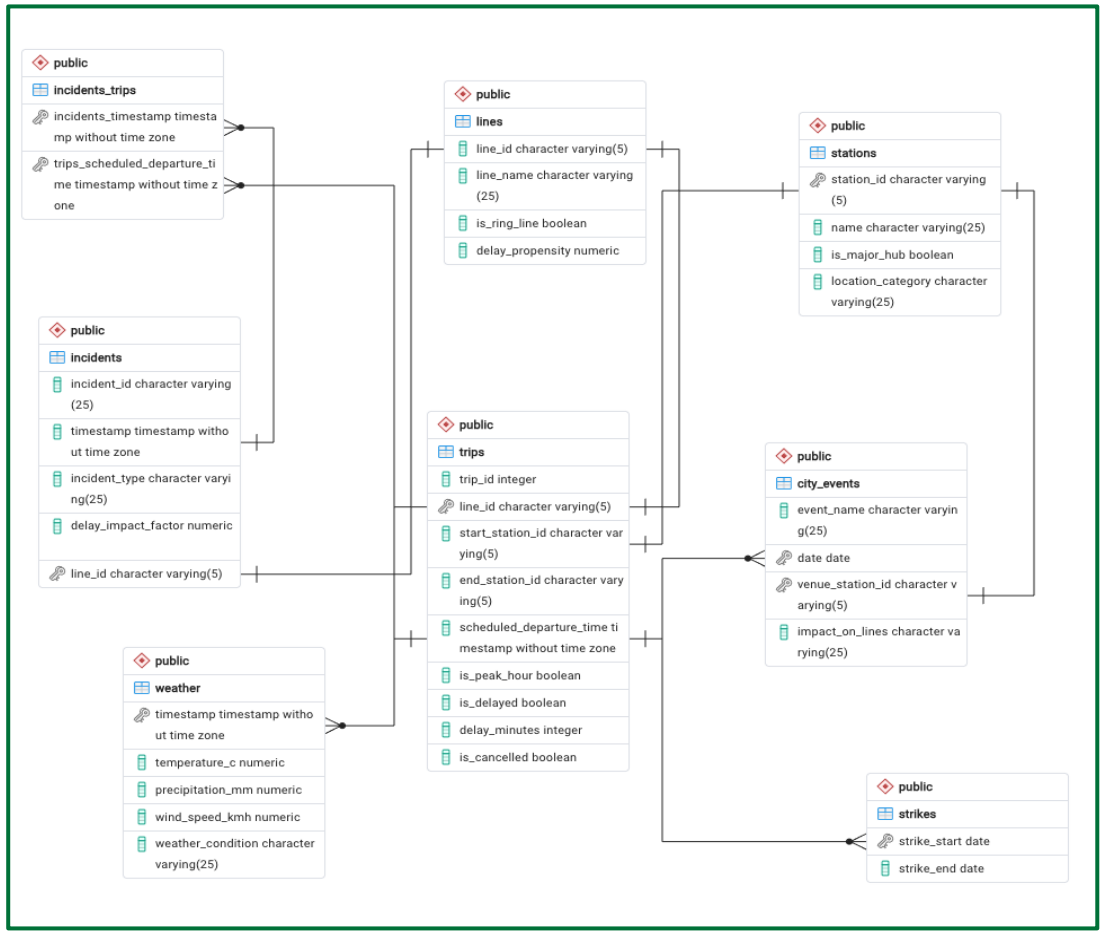
Multi-table relational dataset:

Relational database with 7 interconnected tables – all cleaned and checked for missing values and duplicates.

A creation of an Entity Relationship Diagram (ERD) on PostgreSQL clarified the connections between the tables. Some of the tables, such as strikes and weather, had to be joined to the trips table through timestamps.

Table Statistics:

131, 771 trips for the year	8,761 hourly weather records	
36 recorded incidents	6 lines in the network	
10 main stations	6 recorded union strikes	3 large city events



Entity Relationship Diagram

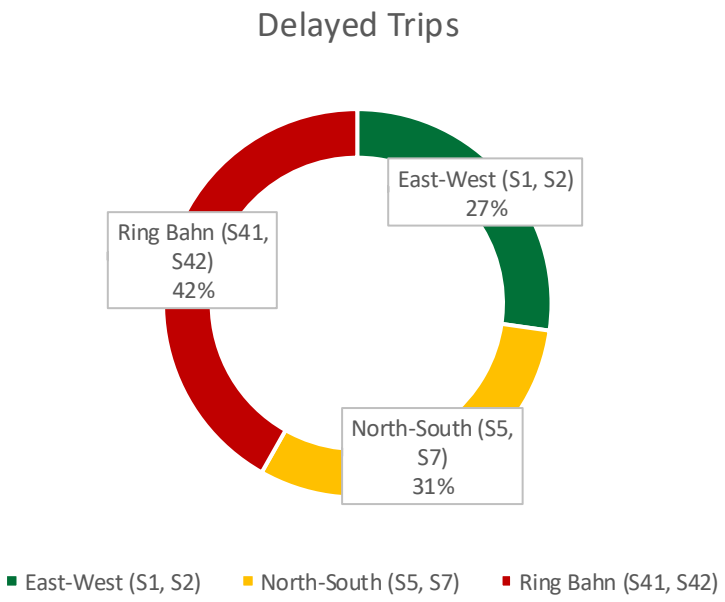
Key Statistics on Delays:

Average delay: 2.81 min	Median delay: 1 min	4% trips marked ‘delayed’ (> 5 min)

Delays by Lines and Incidents (Infrastructure)

Lines:

The S41 and S42 (Ring Bahn) have the highest rates of delayed trips (82%) and cancellations across all 6 lines.



The Ring Bahn's circular route serves dense residential areas with more stops and transfer points, increasing exposure to disruptions.

Incidents:

Power outages carry the highest incident impact factor (**9.88**), far above signal failures (4.73) and track maintenance (2.47).

A correlation test is run between delay impact factor and delay minutes. The CTE defines the time frame between one hour before an incident and one hour after the incident.

Query: Correlation between Incident Severity and Delay Duration

```sql
WITH inc AS (SELECT A.line_id, A.incident_type, A.delay_impact_factor AS dif,
    A.timestamp, ROUND(AVG(B.delay_minutes), 2) AS adm
FROM incidents A
JOIN trips B ON A.line_id = B.line_id
WHERE B.scheduled_departure_time BETWEEN A.timestamp - interval '1 hour'
    AND A.timestamp + interval '1 hour'
GROUP BY A.line_id, A.incident_type, A.delay_impact_factor, A.timestamp)

SELECT
    (COUNT(*) * SUM(dif * adm) - SUM(dif) * SUM(adm)) /
    (SQRT(COUNT(*) * SUM(POWER(dif, 2)) - POWER(SUM(dif), 2)) *
    SQRT(COUNT(*) * SUM(POWER(adm, 2)) - POWER(SUM(adm), 2)))
    AS correlation_coefficient
FROM inc;
```

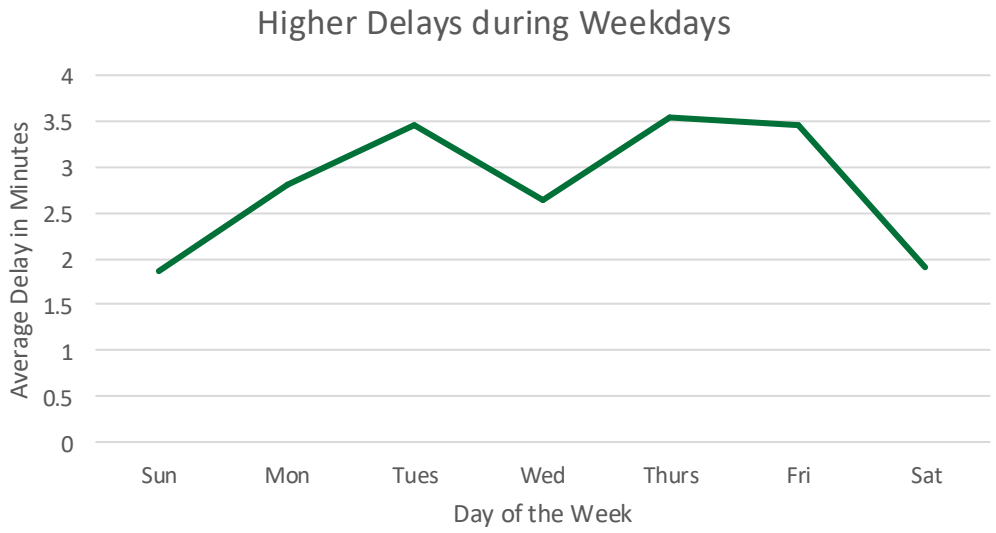
Result: -0.11 = No correlation, suggesting the impact factor is not measured by delay duration.

Delays by Hours, Days and Months (Time)

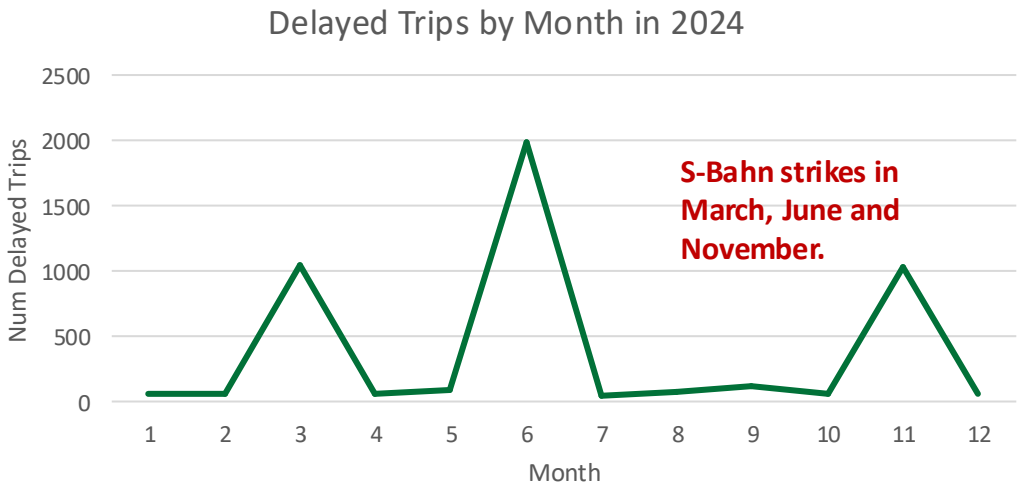
Peak hours (6:00–10:00 and 16:00–20:00) show a 92% delay rate versus 75% during off-peak hours.

Peak-hour delays are mostly 1–5 minutes — consistent with minor congestion from high traffic volume rather than major incidents.

Weekdays have higher average delays than weekends; commuter traffic is the main driver of minor delays



Three months out of the year 2024 had significantly more delayed trips. These months coincide with major strikes by the GDL, a German trade union for train drivers.



How disruptive were the strikes on overall rates of delay and cancellation?

Forceful Impact of Strikes

Strikes are the biggest disruptor: average delay jumps from 1.5 minutes to 22 minutes on strike days, with a 10% cancellation rate.

	Regular Days	Strike Days
Average Delay	1.52 min	22 min
Maximum Delay	13 min	1181 min
Cancellation Rate	0.65 %	10 %

Further investigation of outliers show that **90 %** of all trips fall within the upper bound of 3.5 minutes.

Query: Identification of Outliers with the IQR Method

```
WITH outlier_delay AS (SELECT *
FROM trips
WHERE delay_minutes > (SELECT percentile_cont(0.75) WITHIN GROUP(ORDER BY delay_minutes)
    +(1.5*(percentile_cont(0.75) WITHIN GROUP (ORDER BY delay_minutes) - percentile_cont(0.25)
    WITHIN GROUP (ORDER BY delay_minutes))) FROM trips))

SELECT delay_minutes, COUNT(*)
FROM outlier_delay
GROUP BY delay_minutes
ORDER BY delay_minutes;
```

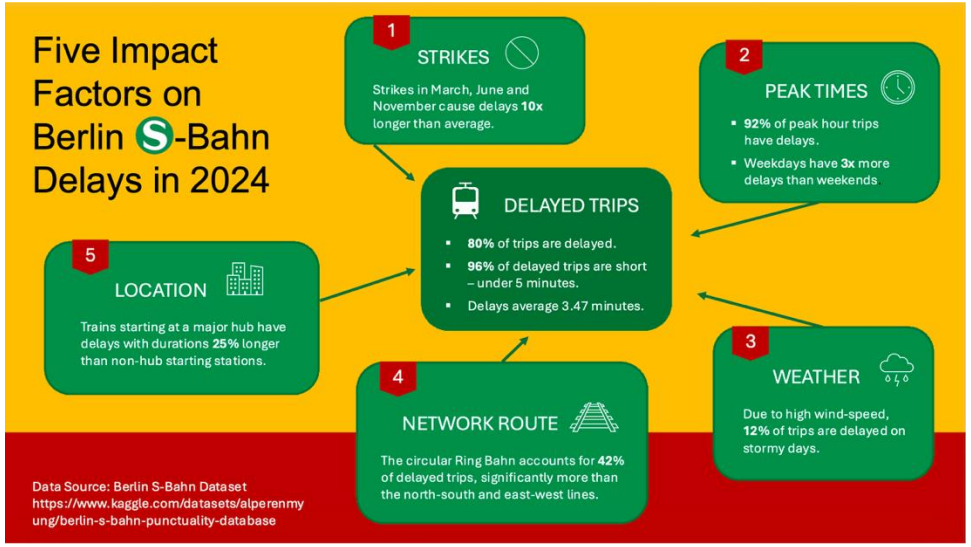
A low number of extremely high delays on strike days is distorting and skewing the dataset to the right.

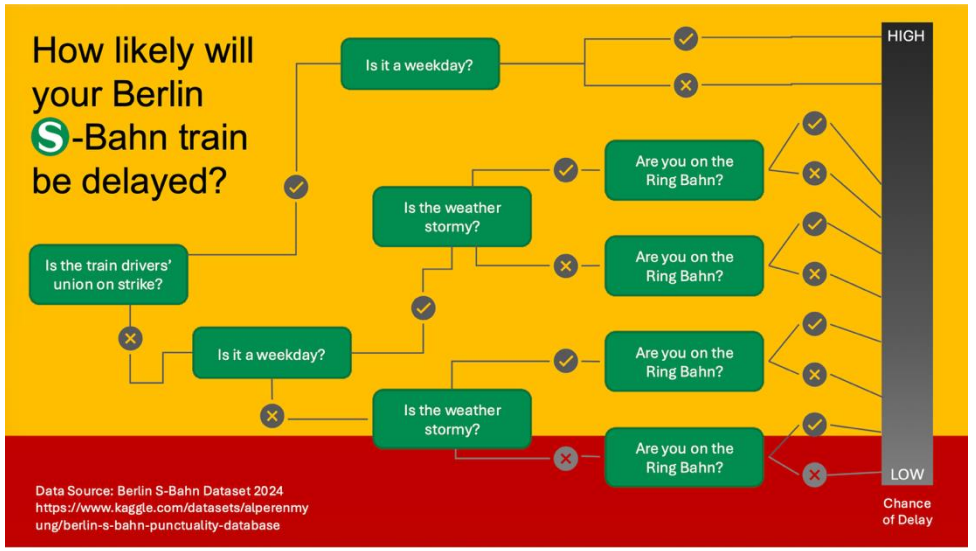
Removing strike days cuts average delay by more than half and reduces the maximum delay from 1,181 minutes to just 13 minutes.

Summary and Deliverables

Key Insights:

- Strikes are the single biggest disruptor — causing extreme delays and a high cancellation rate.
- Peak hours are the primary driver of minor delays (1–5 min), pointing to commuter traffic as the root cause.
- The Ring Bahn (S41/S42) is the most delay-prone line group, with the highest rates of delayed trips and cancellations.





The full results of the analysis were delivered in two formats:

- An infographic showing the five impact factors on S-Bahn delays.
- A decision tree diagram (*interactive on PowerPoint*) showing hierarchy of impacts.

Conclusion

Project Assessment:

This project gave me hands-on practice querying a multi-table relational dataset in PostgreSQL, including joins, subqueries, and aggregation.

The most significant analytical decision was separating strike days from the rest of the data — this transformed the findings and produced a much more accurate picture of everyday S-Bahn performance.

This analysis also offered me experience in managing an end-to-end project: from inputting the data into PostgreSQL and forming an ERD, to cleaning data and querying for answers, and finally to summarizing findings in an Excel report as well as visual presentations on PowerPoint.

Next Steps:

Predicting delays: query a merged dataset and apply machine learning methods to explore whether delays can be predicted based on the various impact factors.

A full analysis workbook and query log are available on request and on GitHub:

ClimateWins

Implementing Machine Learning for Weather Prediction

Case Study

Objective:

ClimateWins, a European nonprofit organization, is interested in using machine learning to help predict the consequences of climate change around Europe and, potentially, the world.

Role:

As a data analyst for ClimateWins, I will assess the tools available to categorize and predict the weather in Europe.

Tools: Python, Jupyter

Data set:

- Observations from 18 weather stations across Europe, dating from late 1800's to 2022
- Values such as temperature, wind speed, snow, precipitation, global radiation and more
- Collected by the European Climate Assessment & Data Set project

Hypothesis:

Machine-learning algorithms can be applied to the data to predict weather conditions.

Process for exploring hypothesis:

- Clean and prepare weather data
- Run optimization algorithm
- Apply various machine learning models, both supervised and unsupervised
- Evaluate accuracy and usefulness of different models on the weather data set
- Propose an effective method for machine learning application for weather prediction

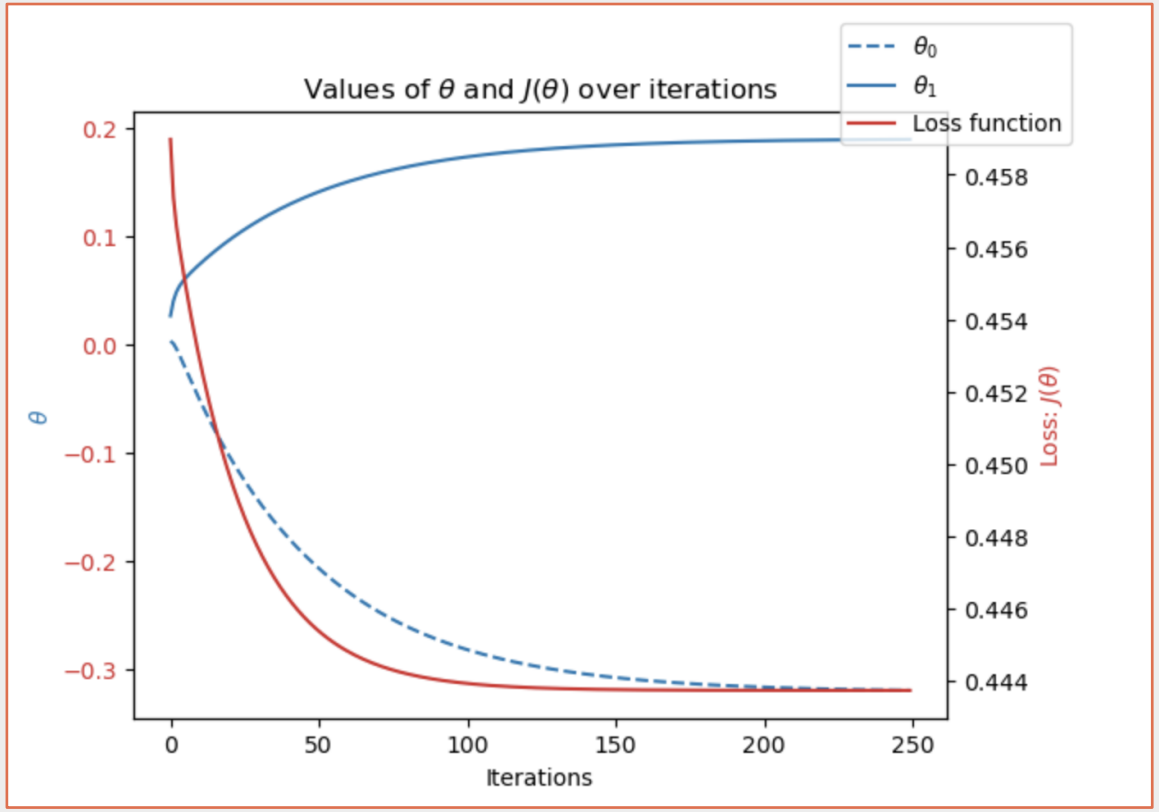
Optimization

To better understand the structure of our data and optimize it for machine learning, we can run an optimization algorithm such as **gradient descent** on a feature of the data.

Here we are fitting a regression model by running gradient descent on daily mean temperatures of various stations for a chosen year.

Gradient Descent on Kassel Station, 1960:

- Iteration: 200, Step size: 0.1
- Successfully converges
- Minimum achievable loss at 0.43



Kassel 1960 Loss Function

Supervised Learning

Preparation:

- An **answers** data set is provided to train the model on predicting if a certain day is pleasant or not.
- The data set is **scaled** to prevent the machine learning model attributing more weight to higher values.
- The data is split, **70%** for training, **30%** for testing.
- For this project, we applied the following models:
  - **K-Nearest Neighbor (KNN)**
  - **Decision Tree**
  - **Artificial Neural Networks (ANN)**

Challenges:

- **Parameters:** For each model, we must experiment on the parameters to produce the highest accuracy.
- **Overfitting:** The models are overfitting to one weather station (Sonnblick), therefore negatively affecting the overall accuracy.

Results on Supervised Models:

Model	Test Accuracy (Confusion Matrix)
KNN	90%
Decision Tree	95%
ANN	95%

Why the **ANN model** works best in predicting current data:

- Most accurate predictions on both training and testing sets.
- Less overfitting than the decision tree model.
- Room for improvement and experimentation with the parameters.

Unsupervised Learning

Neural Network Models

The success of the ANN model in the supervised learning exercise leads to an exploration of the **Recurrent Neural Network (RNN)** model in unsupervised learning. RNN's strength lies in handling **temporal data**.

The **Long Short-Term Memory (LSTM)** model is an improved version of RNN. The model is trained to predict the same weather conditions: pleasant or unpleasant.
Here is the final Keras model layout for running LSTM on the weather data.

```
[52]: epochs = 30
      batch_size =64
      n_hidden = 32

      timesteps = len(X_train[0])
      input_dim = len(X_train[0][0])
      n_classes = len(y_train[0])

      model = Sequential([
          Input(shape=(timesteps, input_dim)),
          LSTM(n_hidden),
          Dropout(0.5),
          Dense(n_classes, activation='sigmoid') ])

[53]: model.compile(loss='categorical_crossentropy',
                    optimizer='rmsprop',
                    metrics=['accuracy'])

[54]: model.fit(X_train,
                y_train,
                batch_size=batch_size,
                validation_data=(X_test, y_test),
                epochs=epochs)
```

Challenge:
It is time-consuming and difficult to experiment with the endless variations of hyperparameters in neural network models. How do we efficiently find the optimal values to improve the performance of deep learning models?

Bayesian Optimization:
Applying a Bayesian search on the hyperparameters produced a set of optimal values for running the model. It improved the accuracy of the pre-optimized model by 5%.

Proposal for ClimateWins

The following proposal results from thought experiments on the possibilities of implementing machine learning to predicting data and the study of various machine learning models throughout this project.

Thought Experiment	Action	Data Required	ML Models
Can machine learning models accurately identify future extreme weather events based on weather data from the last 60 years?	Optimize a model to accurately identify extreme weather conditions.	○ ClimateWins dataset ○ 'Answers' dataset of past extreme weather events	ANN
If we determine that extreme weather events are increasing, can we predict when and how frequent these events will occur?	Train model to predict the time and frequency of weather events.	○ ClimateWins dataset ○ 'Answers' dataset of past extreme weather events	LSTM
If we can predict when and how often these disasters will happen, will we be better prepared? Will there be enough time to make changes?	Analyze risk by location and population to determine the optimal time frame for disaster preparation.	○ ClimateWins dataset ○ Population dataset ○ Location dataset of disaster predictions	LSTM

Conclusion

Key Insights:

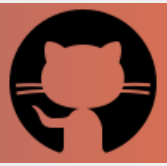
- Machine learning models are able to help predict weather conditions.
- Neural network models work best for the ClimateWins dataset.
- Complex deep learning models such as LSTM will need optimization algorithms to produce higher quality results.
- Thought experiments can assist in creating a viable method for applying machine learning for extreme weather prediction.

Project assessment:

Working with machine learning algorithms is a challenging yet exciting venture. This project has given me an insightful first look at the complex inner workings of various machine learning models. I've gained an understanding of which models are useful in different scenarios, or even that there are possibilities out there still to be explored.

As powerful as machine learning models are, I have also realized through this project how vital the quality and preparation of the data is to running a successful and useful model. In future projects, I would aim to always thoroughly clean and transform the data before modeling.

The project is available on GitHub:



CONTACT




If you have questions, or are interested in working with me, please get in touch!