

Mini-Project: SQL - From Data to Insight

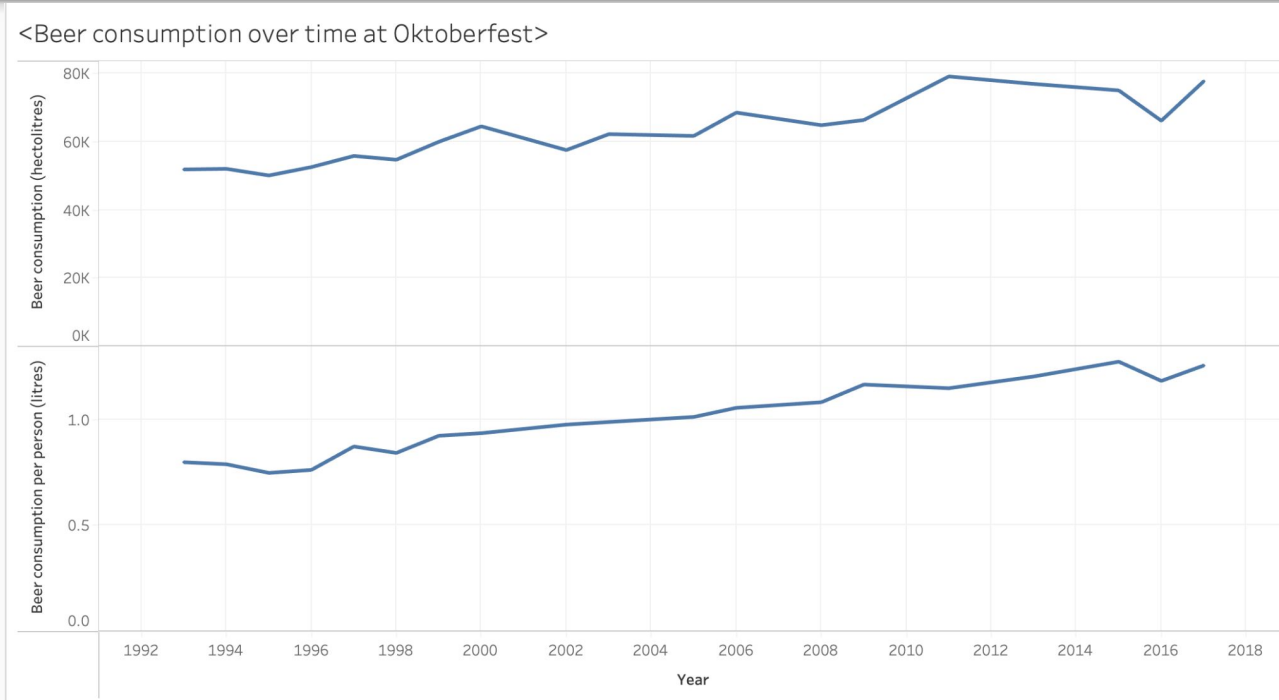
Bart and Matthew

22.01.24-26.01.24 (DA Week 3)

We'd like to understand **what drives beer consumption at Oktoberfest**



General trend: beer consumption at Oktoberfest is increasing



Source: [Kaggle](#)

Which variables might drive beer consumption at Oktoberfest?

FESTIVAL FACTORS

- Price of beer
- Number of visitors
- Chicken consumption

ECONOMIC FACTORS

- Inflation rate in Germany


CULTURAL

- How FC Bayern are performing (as measured by number of goals scored in the month of September)

NATURAL

- Weather (as measured by mean temperature in Munich in September)


We looked at 3 data sources

 PRAJWAL DONGRE · UPDATED 4 MONTHS AGO

◀ 7 New Notebook Download (1006 B)

Octoberfest from 1985-2022

Prost to Oktoberfest: 36 Years of Beer, Chicken, and Revelry ~ 1985 to 2022



Data Card Code (0) Discussion (0)

About Dataset

The **Oktoberfest** in **Munich** (Wien in dialect) is the **largest folk festival** in the world. It has been taking place on the **Theresienwiese** in the **Bavarian capital Munich** since 1810 and is visited by around **six million people** every year.

Source - <https://opendata.muenchen.de/dataset/oktoberfest/resource/e0f064df-6d99-4743-b22b-81a8b18dd1d2>

Usability 8.82

License Other (specified in description)

Expected update frequency Annually

Bundesliga Results 1993-2018

Including half time and full time scores



Data Card Code (3) Discussion (1)

About Dataset

This dataset contains results from every Bundesliga match from 1993-1994 to 2017-2018.

It also includes half time results, but only from 1995-96 to 2017-18. Columns include Division (denoted as D1), HomeTeam, AwayTeam, FTHG (final time home goals), FTAG (final time away goals), FTR (full time result), HTHG (half time home goals), HTAG (half time away goals), HTR (half time result), and season.

This was inspired by the lack of smaller, league specific datasets, in the face of large, all Europe match result sets.

Data compiled into one file from [this site](#), a football betting site from the UK. It contains individual datasets for each season, but I combined them into one single set for ease of use.

Usability 7.69

License Unknown

Expected update frequency Not specified

Tags Football Europe

Startseite → Suche → Tabellenaufbau → Ergebnis

Tabelle

Download: Optionen:

Verbraucherpreisindex: Deutschland, Jahre

Verbraucherpreisindex für Deutschland			
Jahr	2020=100	Veränderung zum Vorjahr	in (%)
1993	67,9	4,5	
1994	69,7	2,7	
1995	71,0	1,9	
1996	72,0	1,4	
1997	73,4	1,9	

[Source: Kaggle](#)



1985-2022

[Source: Kaggle](#)



1993-2018



Required filtering

[Source: genesis-destatis](#)



1985-2022



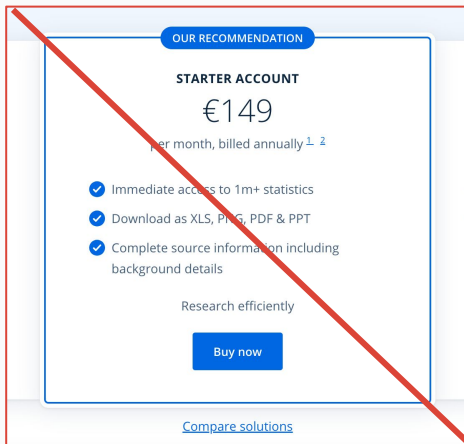
We wanted to look at a 4th data source

Mean temperature in {Munich} or {Germany} by {month} or {year}

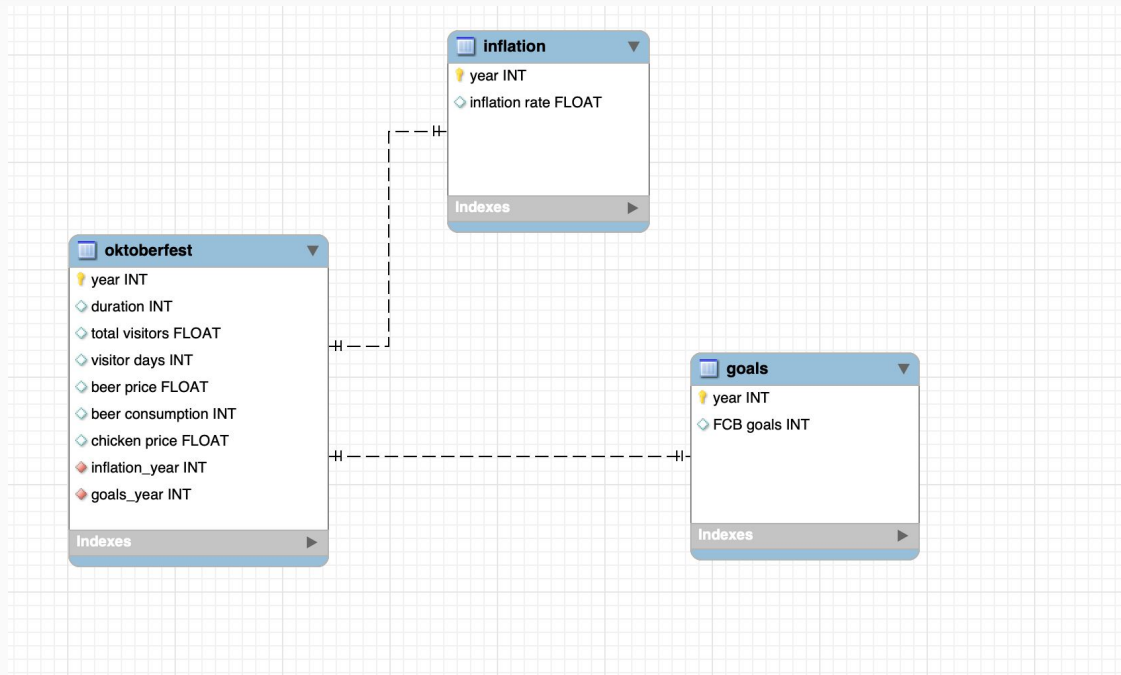
BLOCKER: Unable to find sufficiently comprehensive data for free

TAKE-AWAY:

- Seemingly simple data-sets can be hard to come by
- Can be time-consuming searching for the right data-set



From our 3 datasets, we built a schema




[Source material](#)

We forward engineered schema (*mydb*) to MySQL workbench





We wanted to populate schema from CSVs

 **stackoverflow**

AboutProductsFor Teams

[Log in](#)

[Sign up](#)

[Home](#)

[Questions](#)

[Tags](#)

[Users](#)

[Companies](#)

COLLECTIVES [+](#)

[Explore Collectives](#)

LABS [?](#)

[Discussions](#)

TEAMS

[Stack Overflow for Teams – Start collaborating and](#)

In MySQL Workbench, using "Table Data Import Wizard" to import CSV creates empty table

Asked 8 years, 3 months agoModified 1 year, 6 months agoViewed 67k times

12

I am attempting to import a csv file into a MySQL table using the Table Data Import Wizard. The sample section at the bottom of the Configure Import Settings screen looks fine and when I run the import, it says all of my entries were loaded successfully. However, when I go to view the contents of the table, only the columns are there and none of my actual data loaded. Does anyone know why this might be happening and how to correct it?

EDIT:

These are a few lines from my CSV file:

```
STATION,STATION_NAME,ELEVATION,LATITUDE,LONGITUDE,DATE,MLY-TAVG-NORMAL,MLY-TMAX-NORMAL,MLY-TMIN-NORMAL,Average Temp,Max Temp,Min Temp
GHCND:USW00094085,PIERRE 24 S SD
US,647.4,44.0194,-100.353,201001,218,322,113,21.8,32.2,11.3
GHCND:USW00094085,PIERRE 24 S SD
US,647.4,44.0194,-100.353,201002,246,354,137,24.6,35.4,13.7
```

Featured on Meta

- Updates to the Acceptable Use Policy (AUP) – January 2024
- Site maintenance – Thursday, February 1, 2024 @ 01:00 UTC (Wednesday, January...
- Temporary policy: Generative AI (e.g., ChatGPT) is banned
- January 2024 post from Ryan Polk, Chief Product Officer
- Discussions update: Expansion to all tags in February

Linked

While troubleshooting on SQL, we ran analyses on Jupyter Notebook

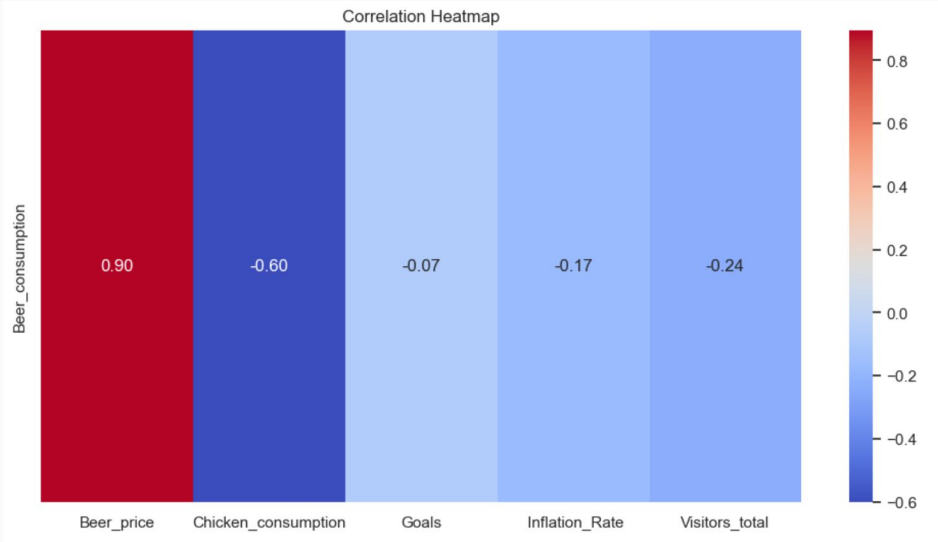
- merged our 3 tables into one 'super-table'

```
merged_df_inner = pd.merge(df1, df2, on='key', how='inner')
```

- ran correlation analysis on our **chosen variables** against **data for beer consumption**

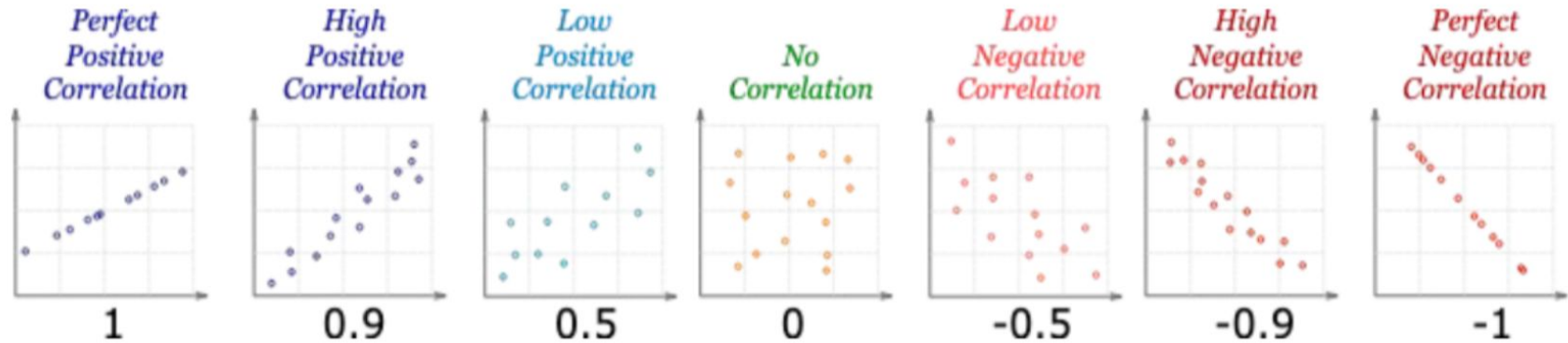
```
correlation_{variable} =  
df['{variable}'].corr(df['Beer_consumption'])
```

While troubleshooting on SQL, we ran analyses on Jupyter Notebook



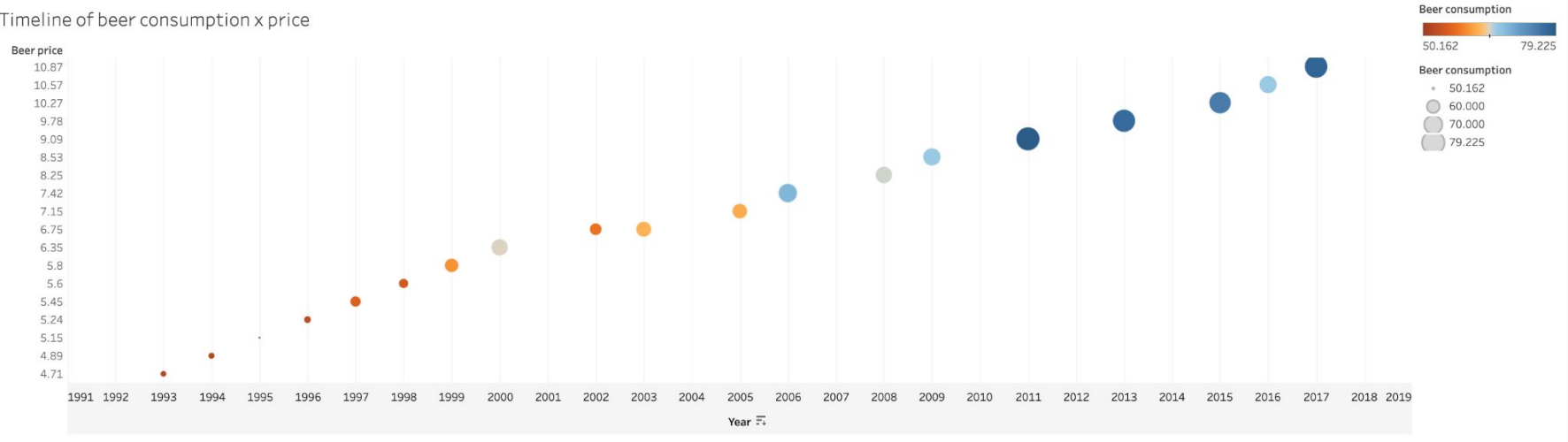
```
1 # VISUALIZATION IN PANDAS
2
3 import seaborn as sns
4 import matplotlib.pyplot as plt
5
6 # Create a DataFrame with correlation values
7 correlation_data = pd.DataFrame({
8     'Beer_price': correlation_beer_price,
9     'Chicken_consumption': correlation_chicken,
10    'Goals': correlation_goals,
11    'Inflation_Rate': correlation_inflation,
12    'Visitors_total': correlation_visitors
13 }, index=['Beer_consumption'])
14
15 # Create a heatmap using seaborn
16 sns.set(style="white")
17 plt.figure(figsize=(12, 6))
18 sns.heatmap(correlation_data, annot=True, cmap='coolwarm', fmt=".2f")
19
20 # Show the plot
21 plt.title("Correlation Heatmap")
22 plt.show()
23
```

Correlation: Guide



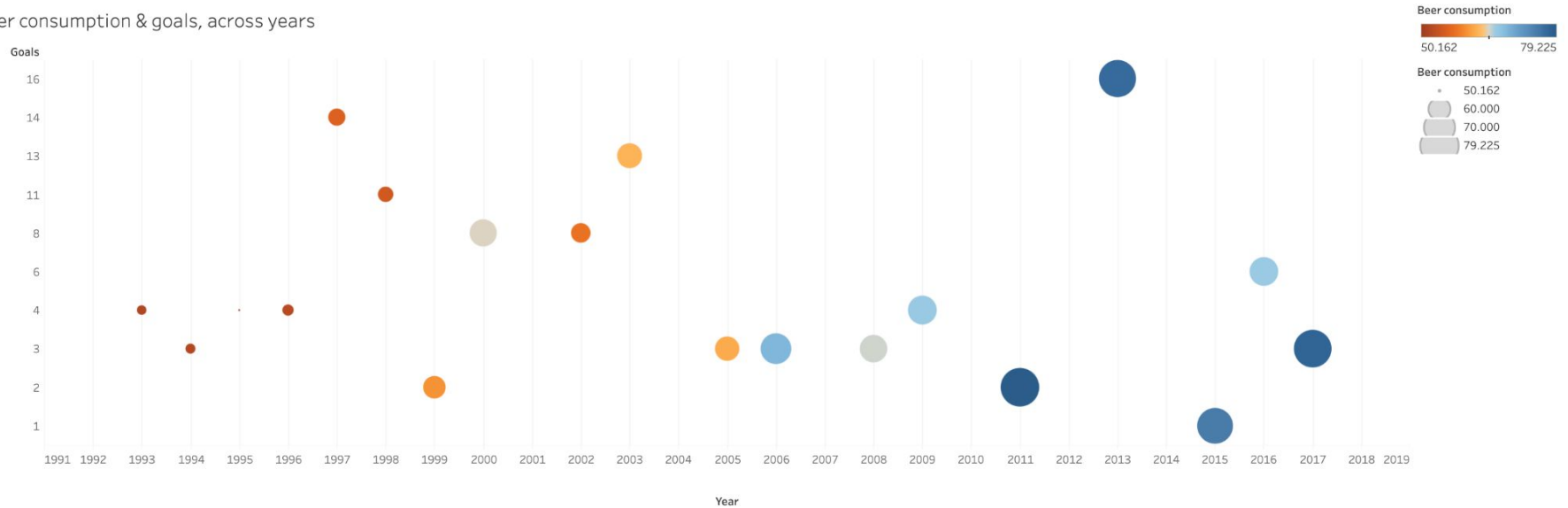
High positive correlation: beer price

Timeline of beer consumption x price



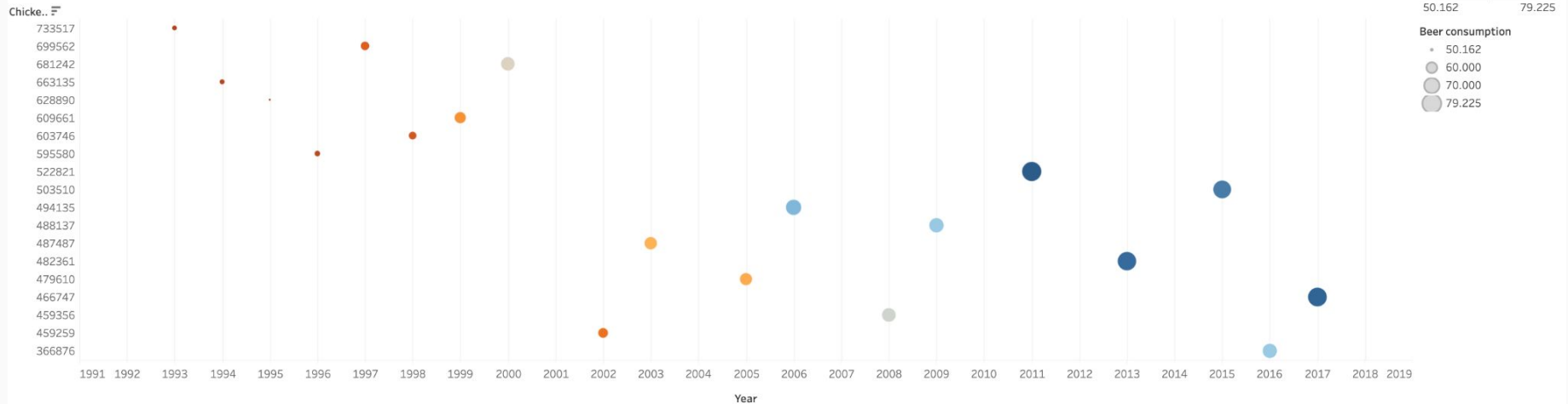
No correlation: FC Bayern goals

Beer consumption & goals, across years



Low negative correlation: chicken consumption

beer consumption chicken across years





Challenges with Tableau

- aligning data types across tools - floats in python became strings in Tableau
 - inhibited further visualisation such as running animation

Conclusion

- only one variable (beer price) correlated to beer consumption (but relationship inverse to that which was expected)
- this investigation would need to draw on more societal variables (e.g. overall drinking habits)
- working with data across multiple tools was more time-consuming than expected; this was to the detriment of our ability to analyse and potentially introduce more variables

Feedback following presentation on 26.01

- would be useful to analyze 'purchase power' as beer price inflation YoY vs inflation YoY; it could be that beer is becoming relatively cheaper, which drives consumption
- schema could be simplified to one table in this instance as there is little value in the additional tables