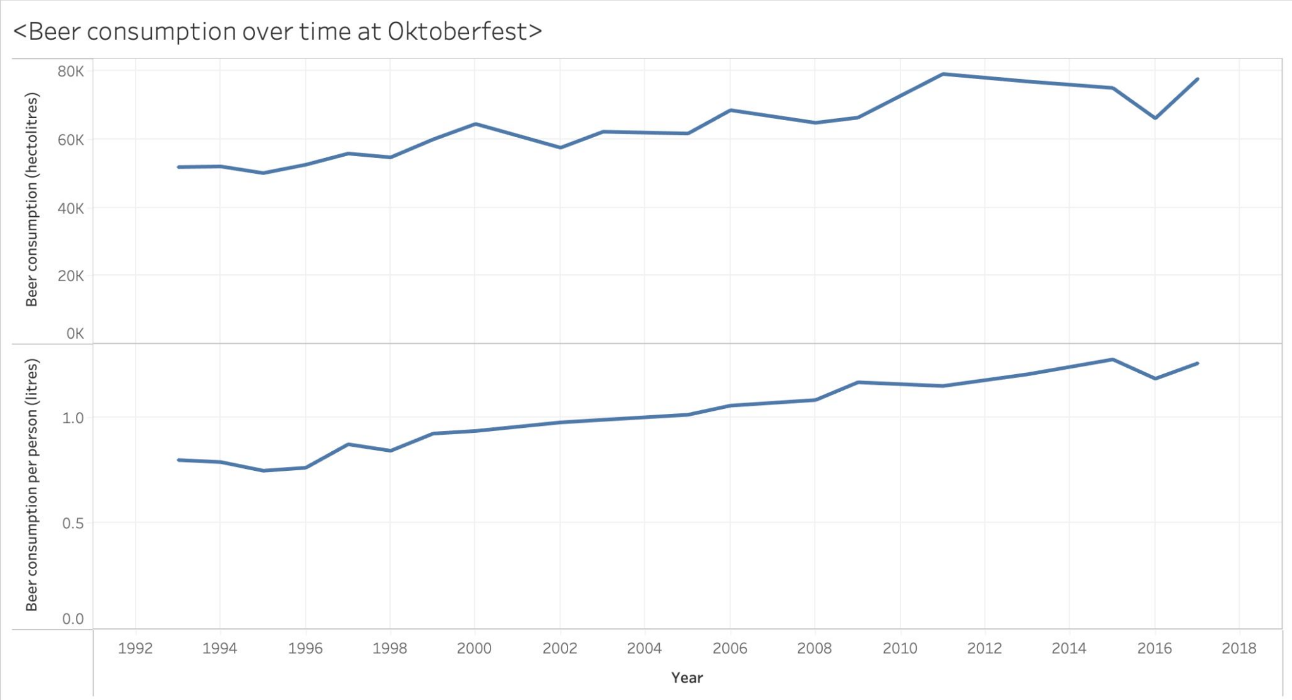
Mini-Project: SQL - From Data to Insight

Bart and Matthew
22.01.24-26.01.24 (DA Week 3)

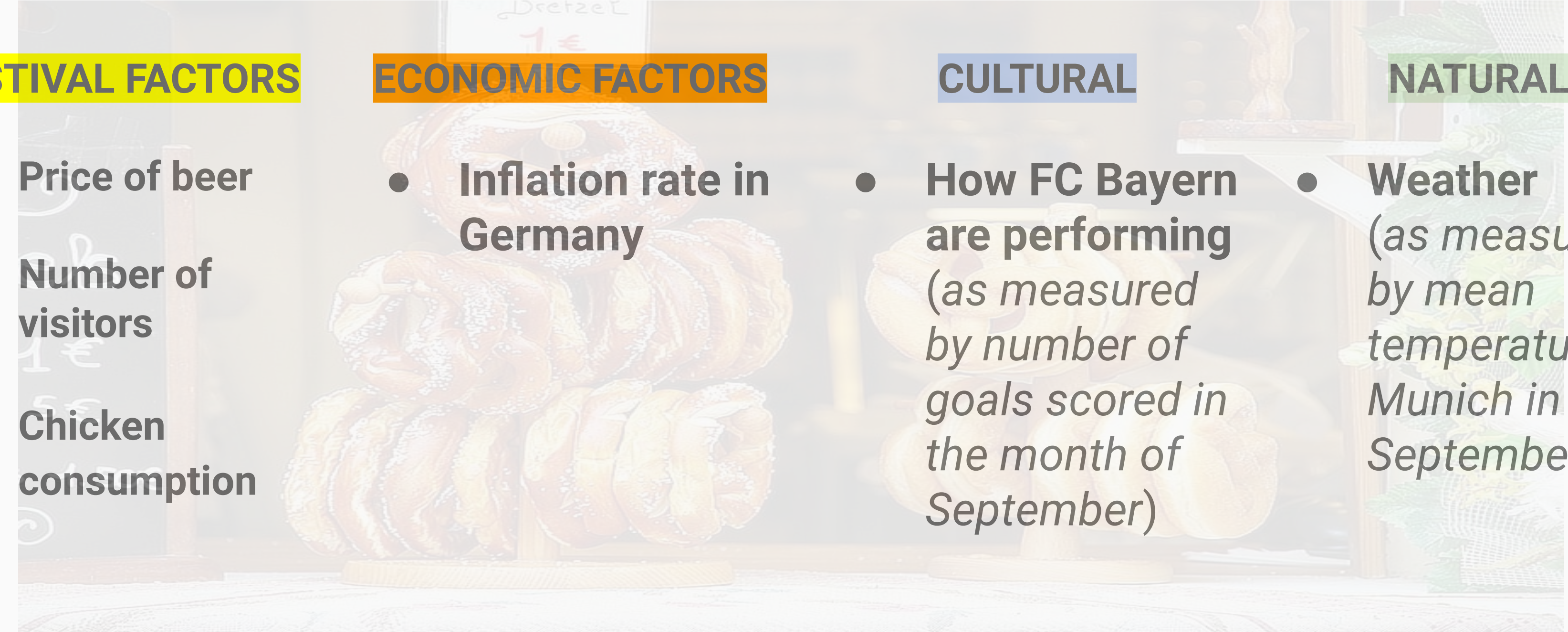
We'd like to understand **what drives beer consumption at Oktoberfest**

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**

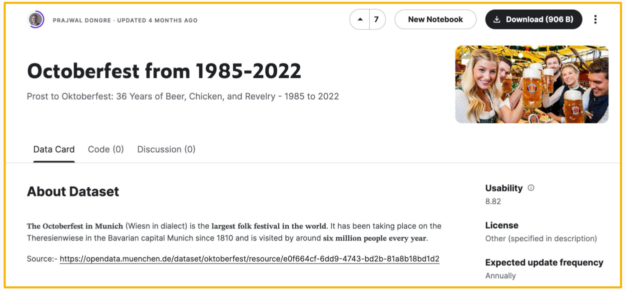
CULTURAL

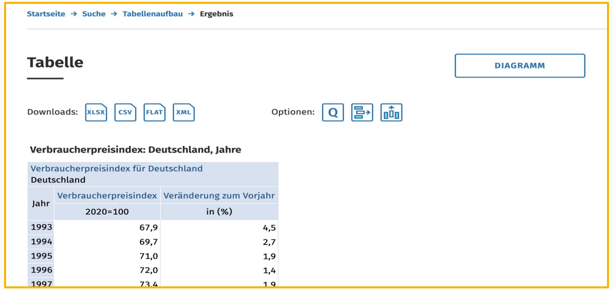
- **How FC Bayern are performing** *(as measured by number of goals scored in the month of September)*

NATURAL

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

We looked at **3** data sources





[Source: Kaggle]

1985-2022

[Source: Kaggle]

1993-2018

Required filtering

[Source: genesis-destatis]

1985-2022

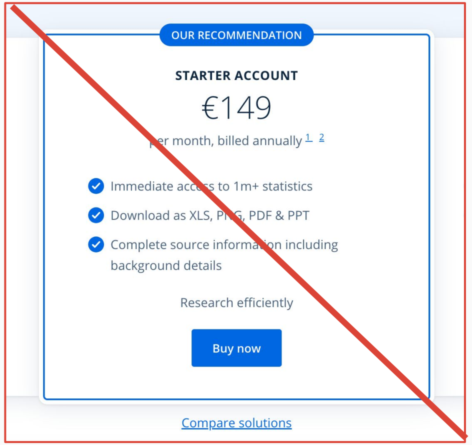
We wanted to look at a 4th data source

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

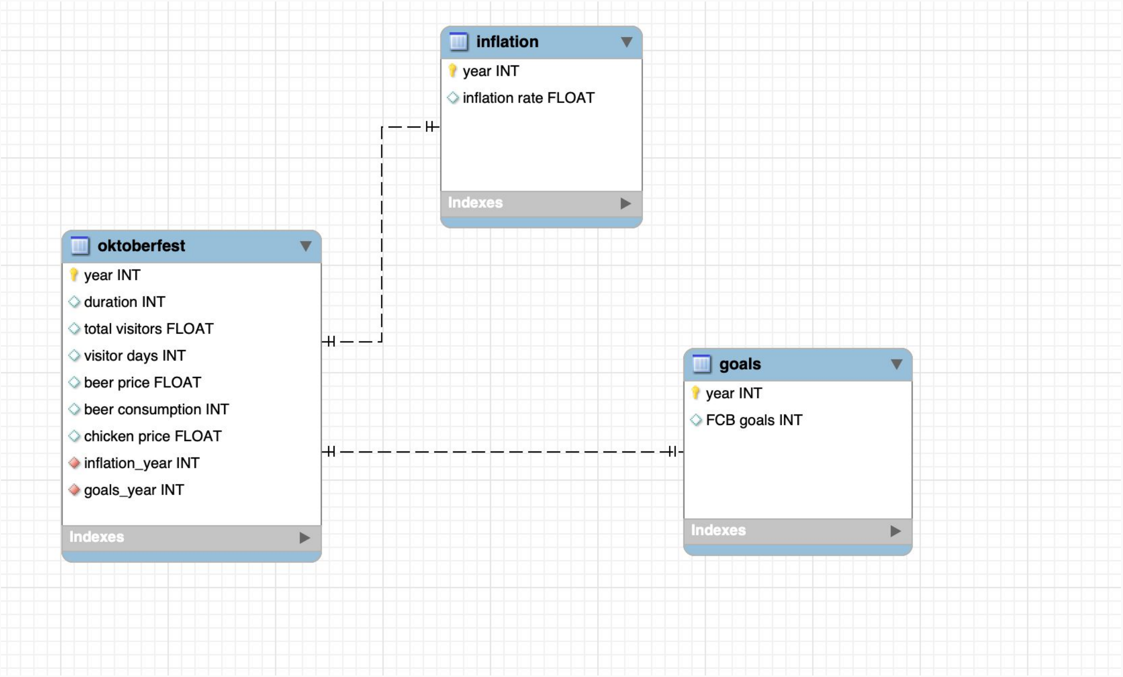
BLOCKER: Unable to find sufficiently comprehensive data for free

TAKE-AWAY:

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



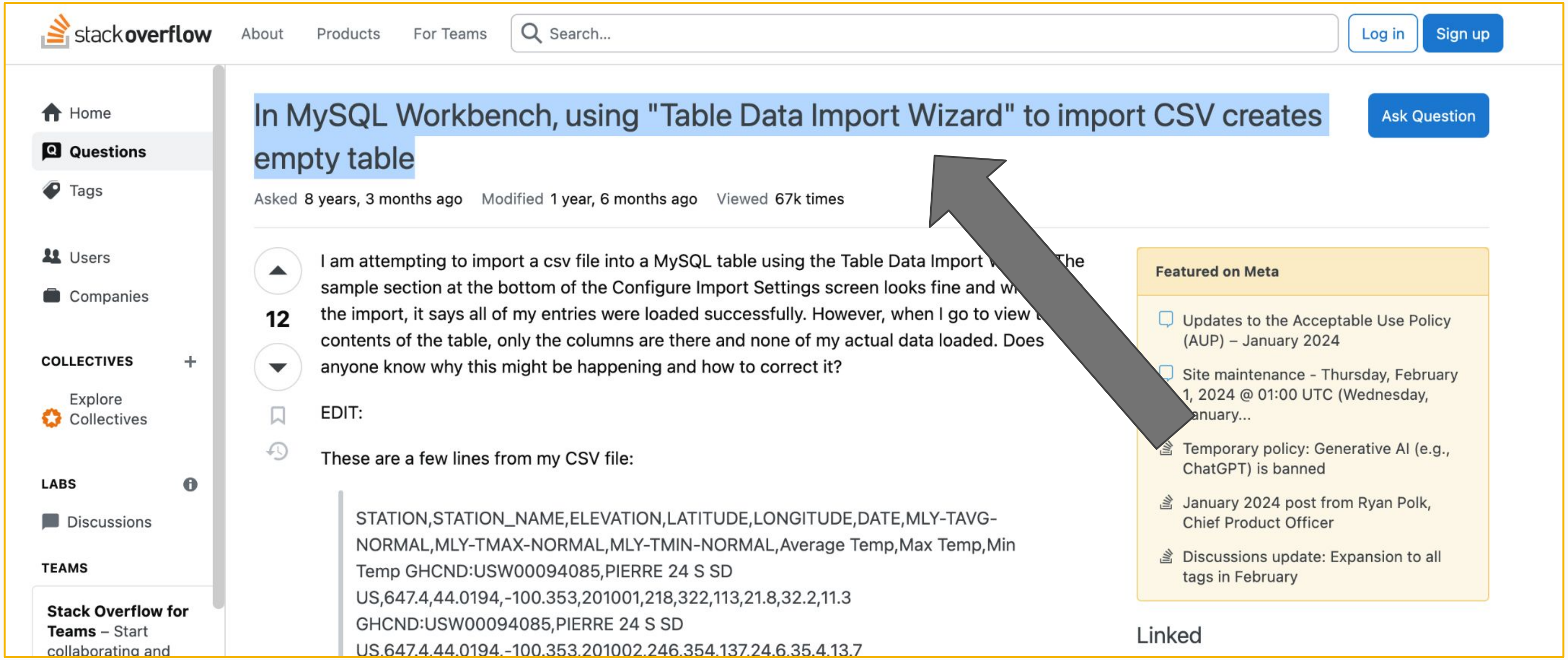
From our 3 datasets, we built a schema



[Source material]

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

We wanted to populate schema from CSVs



While troubleshooting on SQL, **we ran analyses on Jupyter Notebook**

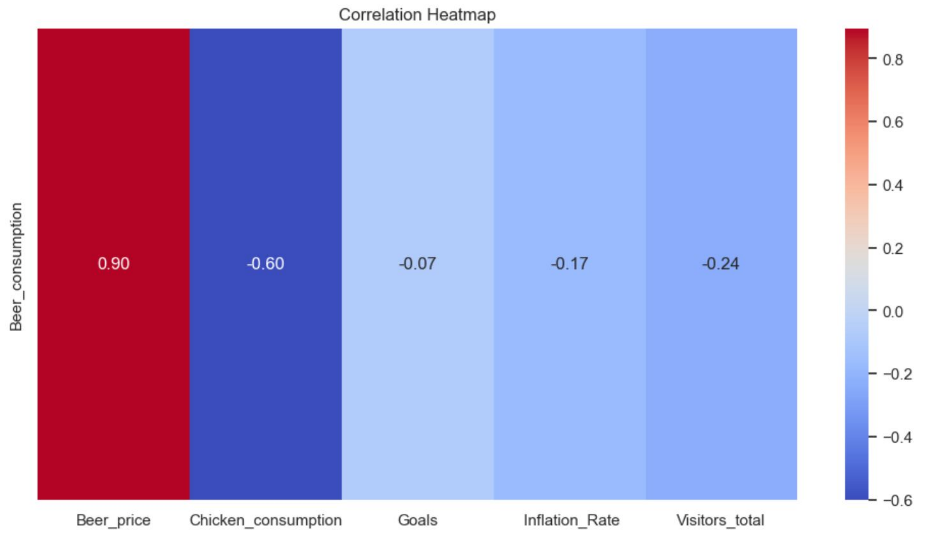
- merged our 3 tables into one ‘super-table’

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

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

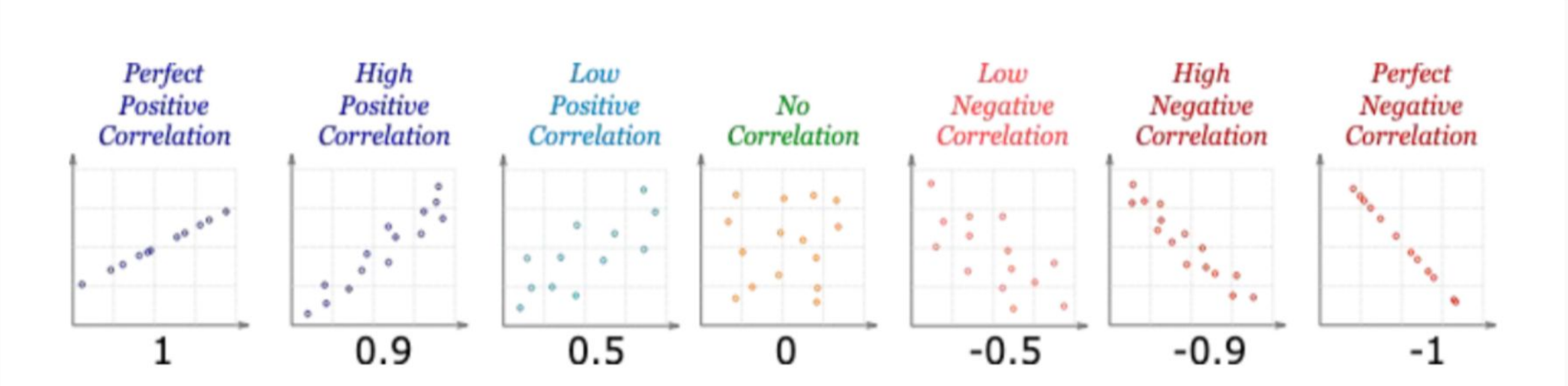
```
correlation_{variable} =
df['{variable}'].corr(df['Beer_consumption'])
```

While troubleshooting on SQL, **we ran analyses on Jupyter Notebook**

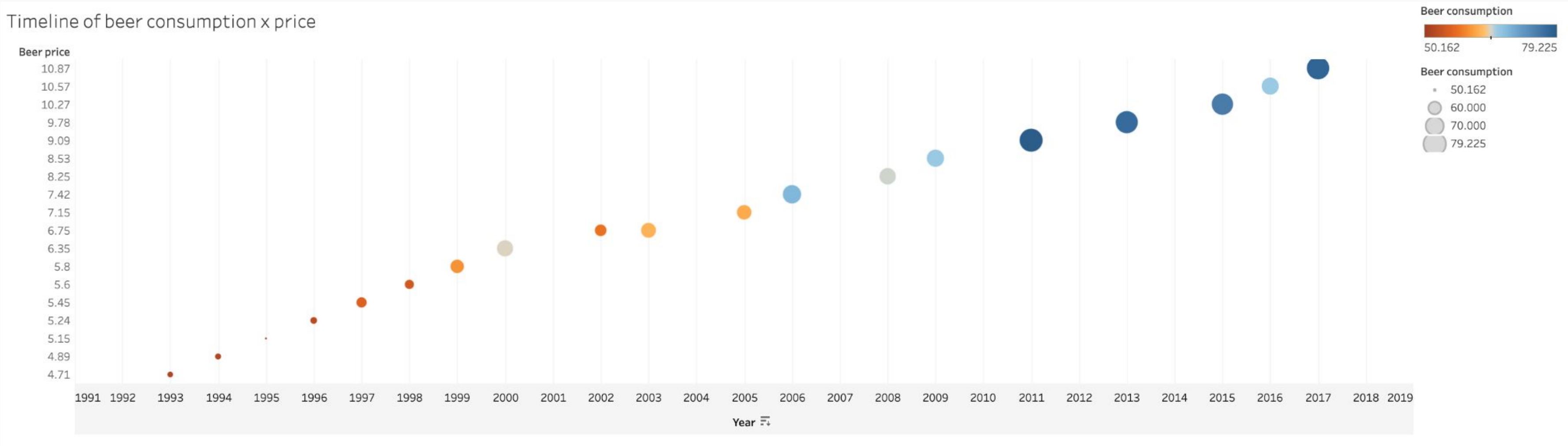



```python
# VISUALIZATION IN PANDAS

import seaborn as sns
import matplotlib.pyplot as plt

# Create a DataFrame with correlation values
correlation_data = pd.DataFrame({
        'Beer_price': correlation_beer_price,
        'Chicken_consumption': correlation_chicken,
        'Goals': correlation_goals,
          ' Inflation_Rate': correlation_inflation,
          'Visitors_total': correlation_visitors
}, index=['Beer_consumption'])

# Create a heatmap using seaborn
sns.set(style="white")
plt.figure(figsize=(12, 6))
sns.heatmap(correlation_data, annot=True, cmap='coolwarm', fmt=".2f")

# Show the plot
plt.title("Correlation Heatmap")
plt.show()

```

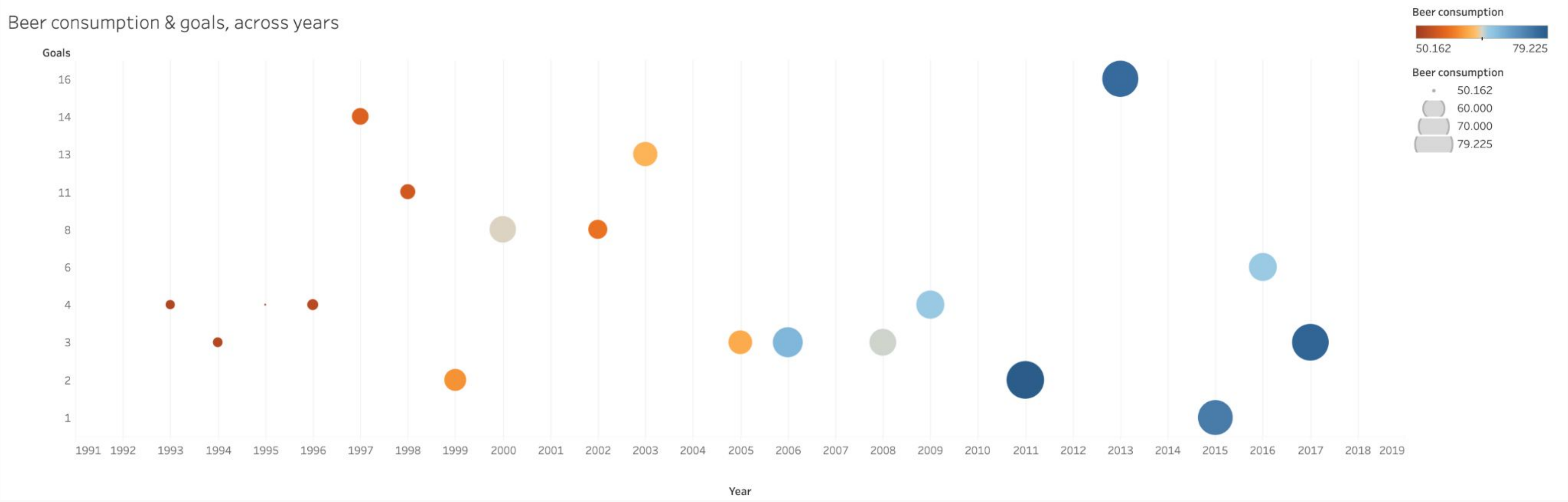
Correlation: Guide



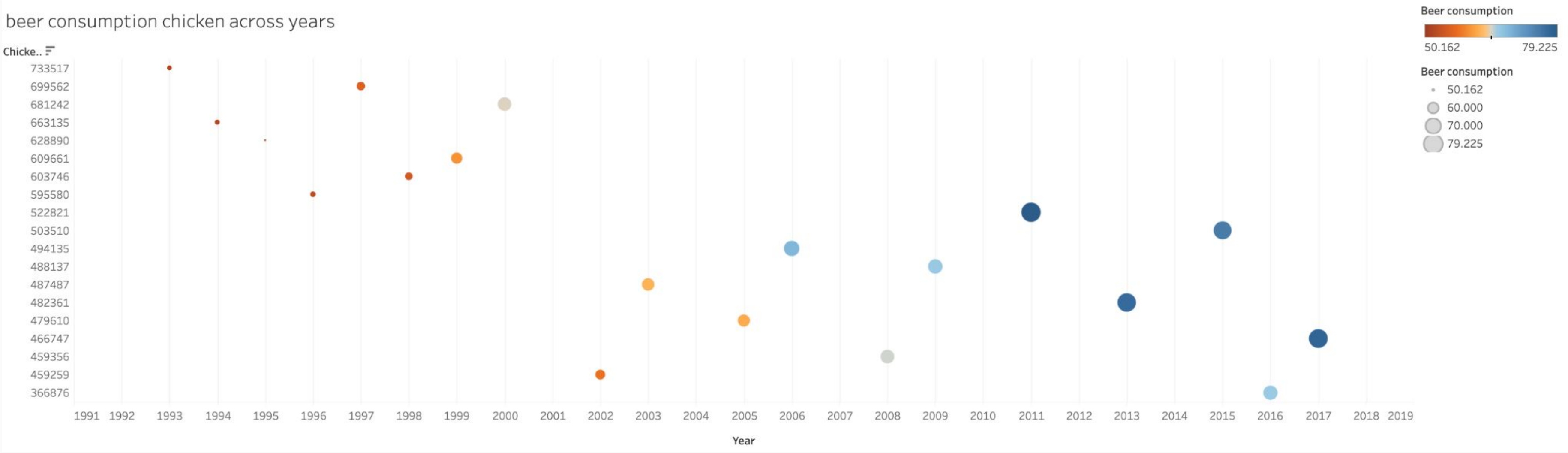
High positive correlation: beer price



No correlation: FC Bayern goals



Low negative correlation: chicken consumption



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