Technische Universität München

Evaluation of Oktoberfest Dataset from 1985 to 2017.

BY

Name	**Matric No**
Saptarshi Mitra	03694168
Sreetama Sarkar	03708506
Sayanta Roychowdhury	03709791
Amusa Oriyomi	03722547

Table of contents

Introduction: **Project description**

In this project, we aim to integrate linear algebra and Python programming to solve real world analytical problems. Such as, predicting the price of beer during the Oktoberfest festival, and also evaluating the total amount of beer that would be consumed during the festival period.
For the remaining part of this report, we would give a brief information on the task, the methodology used in carrying out the task, the result obtained and finally the discussion of the results.

Task
The task we would cover in this project is divided into two sections:

- The first section involves the data analysis of a given set of data matrix A. With the data provided, we would determine the dimensions of the matrix A and evaluate the features that describe the data. Also, we would factorize the matrix A using SVD. The outcome of this will be obtaining the plot of diagonal matrix S, estimation of the rank with singular Values and finally performing a low rank approximation of matrix A.
- The second section of our task involves learning from the data set obtained. We are going to predict the potential beer price and total amount of beer that would be consumed during the festival.

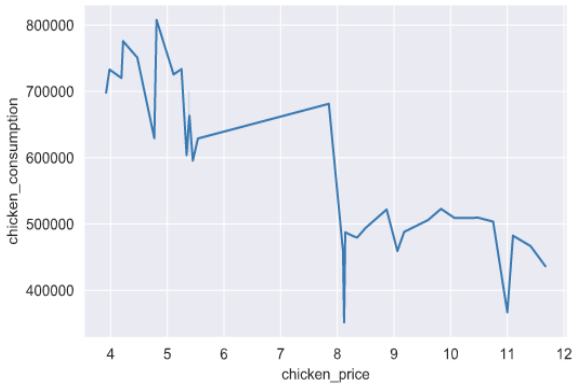
RESULTS
Question 1a
Look into the data matrix A. Which dimension does A have? Which features describe the data? Does the dataset contain null values?
Answer:

1. The shape of data matrix A obtained is: (34, 8)
2. The columns or features that describe the data are: year, duration, visitors total, visitors day, beer price, beer consumption, chicken price and chicken consumption.
3. The dataset does not contain any null values.

Question 1b
Plot with Seaborn various data dependencies (e.g. beer price vs. beer consumption) and depict the most interesting ones.

From the plot, it showed that the number of visitors are random with years, but the year 1988,2001,2009 and 2016 recorded the lowest count.

What is the relationship between beer price and beer consumption? Is the relationship between chicken price and chicken consumption similar? Try to discover other interesting findings. What relationship would you never have expected? What variables are highly correlated? Using Seaborn, plot various data dependencies (e.g. beer price vs. beer consumption) and depict the most interesting ones.

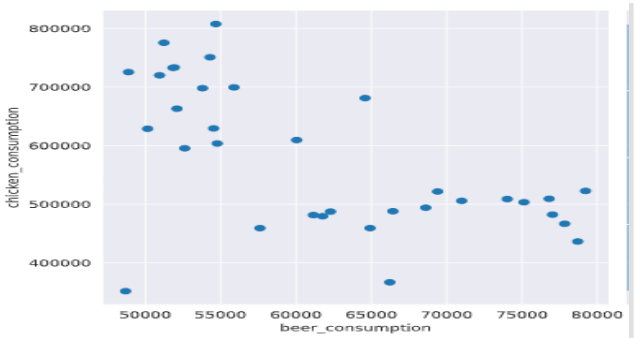


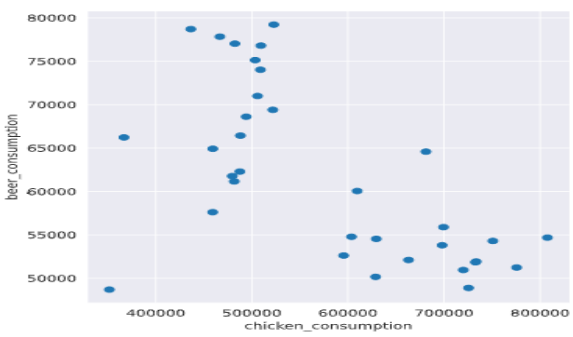
As shown in the graph, Beer consumption increases even with increase in beer price, whereas chicken consumption decreases with the increasing chicken price.

From pair-grid plot,
One interesting observation can be that the visitors/day has a decreasing trend over the years, also chicken consumption decreased with the increase in years. Beer consumption and chicken consumption can be found to be inversely proportional.

From the heatmap,
a) Year and beer price, b) Year and chicken price, c) Beer consumption and year, d) Beer price and beer consumption e) Beer price and chicken consumption are found to be highly correlated (positively or negatively). An increase in beer consumption with the increase in beer price is quite unexpected

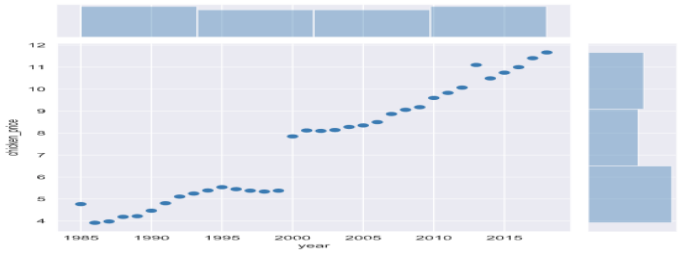
Except one outlier usual trend is that the chicken consumption reduces with higher beer consumption most beer consumption values and chicken consumption values are centered around 50000 in histogram





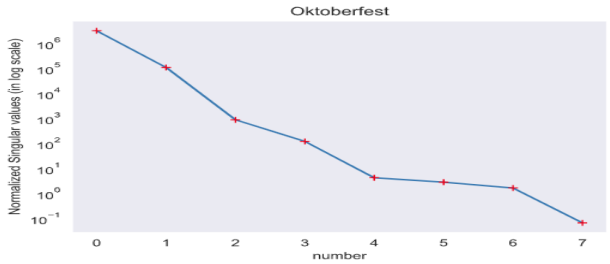
So due to the price increase in the previous year, there was around 2000 less chicken consumption.
.
Question 1c
Let's have a look at the year and chicken price. Use joint plot to plot the relationship between these two variables. What could be the reason for the jump in the chicken price between 1999 and 2000?

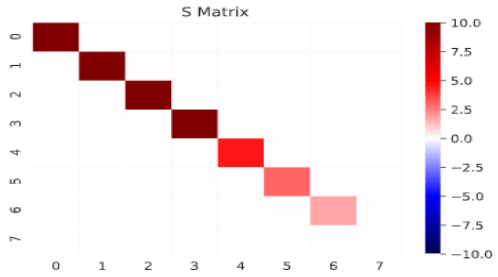
A sharp fall in chicken consumption could be the reason for the jump in chicken price between 1999 and 2000.



Question 1d

Factorize A using the Singular Value Decomposition





We can say that the rank is 8 though the last sigma is near to 0, order of 10^-2. So if low rank approximation is taken rank 7 can also be taken**.**

Question 2a

Now, choose a variable which you want to predict (e.g. predict the beer consumption based on all other features). Solve the least-squares problem for this variable

In this part we predicted the beer consumption based on all other features. The training set consisted of 23 examples whereas the test set had 11 examples. The least square solution was used to predict the beer consumption values. We obtained an average **training error** of **617.62** and average **test error 1292.03** in beer consumption. As expected, the training error was less than the test error. So, we can conclude that the model did not overfit

Question 2c

What is the test error? What is the training error? What deviations do you observe between training and test set error?

We chose to drop 'duration', 'visitors total' and 'visitors day' since from the pair plot these features appear to be uncorrelated with beer consumption. While computing the least square error in this case the average training error was found to be 730.3 and average testing error was 1089.84. So, from the previous case, there is a decrease in test error but an increase in training error. Therefore, we can conclude that dropping these features performs a better generalization to the dataset.

Question 2d

Standardize the data matrix A into a new, normalized dataset B. Compare the results obtained from solving the Least Squares problem with the dataset A (not standardized) with the results obtained from the standardized dataset B.

After normalizing the dataset, the training error obtained was 729.03 and test error was 1047 for beer consumption. We found the results for standardized dataset comparable with the results for the non-normalized one.