Statistical Analysis for the Prediction of Corn Yield in Various US States

Benjamin Leidig, Chavosh Khazeni, Ryan Sponzilli,
Josh Lee, Max Zhang, Max Wong

Project Lead: Danny Silverstein

Data Mining Project

Contents

1 Abstract

2 Introduction

Being a large source of food and raw material, corn is an important crop to the United States. However, there are always fluctuations in corn yield due to external factors such as weather or temperature. Now that climate change is accelarating, it is becoming harder for scientists to make predictions. Therefore, being able to predict the amount of corn that can be harvested in one year is crucial for farmers when they prepare for harvest. In this paper, we will be collecting data on various variables- i.e average temperature, precipitation- from major corn-producing states such as Illinois and Iowa over 100 years and comparing them with corn-yield in their respective time period. We will then be drawing conclusions about correlations between variables and trends in the data.

2.1 Data Sources

3 Method

3.1 Choice of Variables

Temperature is an important factor affecting crop yield as higher temperatures can increase the length of the growing season. On the other hand, it may also decrease the time it takes for corn to mature. Whether it increases or decreases the yield of corn during harvest season may differ case by case, but it is a factor that we must consider. We took average temperature, maximum temperature, and minimum temperature data from 6 different states.

Precipitation is the measure of rainfall in a certain period. Rainfall is important for corn growth as it is a vital source for their respiration process. Again, we took precipitation data from 6 states.

Population Density is the measure of abundance of human density in a given area. It is a good indicator of urbanization in an area which can affect air quality, water quality, soil quality, etc. Therefore, we may want to explore the correlation between population density and the corn yield in that area. We collected data from 6 dates.

Price Received measures the income that farmers receives from corn harvests in a given year. This data can be an indicator of willingness to invest in corn the coming year and their preference to grow corn- farmers may decrease corn planting if the price received is low. We collected data from 6 states.

Applications is the pesticides and fertilizers applied to the corn. Chemicals added to plants can affect the growth of plants and also change external factors (such as insect harm). It will be measured using pounds per acre applied.

3.2 Interpolation for Missing Data

Some of the data sets were not entirely filled so we had to clean the data manually. Linear interpolation was performed to fill in these empty data cells by predicting their value using their neighboring cells. The formula employed is given below:

$$y = y_1 + (x - x_1)\frac{y_2 - y_1}{x_2 - x1} \tag{1}$$

where $(x_1, y_1$ and (x_2, y_2) are known data points.

3.3 Data Cleaning

We used the PANDAS package in Python to organize the data. We extracted the two columns that contained the yield and the variable out of the dataset and then used the **merge** function to combine them back into a two column dataset. Below is the example code using temperature data from the state of Iowa

4 Results

5 Discussion