

**STATISTICS FOR DATA SCIENCE  
UE19CS203**

**ANALYSIS OF CAMERA DATASET**

**TEAM MEMBERS:**

**NAME: SRN:**

1. PHANI KUMAR VEDURUMUDI PES2UG19CS281

2. PAWAN PRASAD P PES2UG19CS280

3. RAHUL S BHAT PES2UG19CS315

4. PRATHEEK PES2UG19CS291

Section: E

***Abstract***

The best way to predict the future is to study past behavior. With the help of data analysis, we are studying about the features of the different models of cameras released from 1994 to 2007.We have gathered information about the various features of the camera models like the **Release date, Max resolution, Low resolution, Effective pixels, Zoom wide, Zoom tele, Normal focus range, Macro focus range, Storage, Weight, Dimensions and Price** of the camera models. We have performed data cleaning by replacing the missing numeric data with mean values of the respective features of the dataset and also we have fixed the types and inconsistent capitalization .We have used various graph visualization techniques to study and analyze our dataset. They include **bar chart, histogram, box plot, line graph and scatter plot. We have used boxplot to filter the unwanted outliers of numeric data from our dataset.** The concept of normalization was used to reduce data redundancy and improve data integrity. Hypothesis testing has helped us in estimating the mean values of the various numeric features of our dataset by choosing random samples from the respective features of the dataset. We also wanted to check the relationship between the various features of our dataset. With the help of these methods of data analysis we were able to study about the variation of the features of the different models of cameras over the years.

***Introduction***

A **camera** is an optical instrument used to record images. At their most basic, cameras are sealed boxes with a small hole (the aperture) that allow light in to capture an image on a light-sensitive surface. We have considered the different problem statement on our dataset like **1.** We have found the growth in the production of camera on every year. **2.** We have considered the price and release date features to fix the outliers. **3.** We have considered the price of camera to see the growth in it. In this way we have used the **camera dataset** to study about the various features of the camera models.

In our project, we have taken a dataset which consists of various camera models with their release dates ranging from the year 1994 to 2007.We have performed various techniques to analyse the dataset like **Data cleaning, Data visualisation, Normalisation, Hypothesis testing and Correlation**. We have cleaned the dataset by relacing the null values with the mean value of the respective columns(features). We have done Data visualisation by plotting various graphs. We have done the Hypothesis testing on three of the features of our dataset. We have also normalised the data by making the mean and variance of each numeric column to 0 and 1 respectively. We have also performed the correlation of each pair of possible features of our dataset.

***Dataset***

Our camera dataset shows us about the different features of the different camera models released from 1994 to 2007.We have taken this dataset from the source: **www.kaggle.com** repository**.** This dataset consists of **1038 observations** and **13 features** which are:

* The Categorical Variables are **Model.**
* The Discrete Numerical Variables are **Max resolution, Low resolution, Effective pixels, Zoom wide, Zoom tele, Normal focus range, Macro focus range, Storage, Weight, Dimensions and Price**.
* The Continuous numerical variables are **Release Date**.

The percentage of missing values in our dataset is **5%**.

**About the features:**

**1.MODEL**:

This feature of the dataset describes about the various models of the different type(company) of cameras present in the chosen dataset.

**2.RELEASE DATE:**

This feature of the dataset describes about the various release dates(in years)of the different models of the cameras in the dataset.

**3.MAX RESOLUTION:**

This feature of the dataset describes about the maximum resolution of the different models of the cameras present in the dataset.

**4.MIN RESOLUTION:**

This feature of the dataset describes about the minimum resolution of the different models of the cameras present in the dataset.

**5.EFFECTIVE PIXELS:**

This feature of the dataset describes about the different number of effective pixels of each camera model in the dataset.

**6.ZOOM WIDE:**

This feature of the dataset describes about the short focal length of the different cameras in the dataset.

**7.ZOOM TELE:**

This feature of the dataset describes about the long focal length of the different cameras in the dataset.

**8.NORMAL FOCUS RANGE**:

This feature of the dataset describes about the lens of the various cameras in the dataset with a **focal** length approximately equal to the diagonal of the film format or of a digital **camera's** image sensor. Most 35mm **cameras normal** lenses have a **focal** length of approximately 50 mm.

**9.MACRO FOCUS RANGE:**

This feature of the dataset describes about the **focal** length of the various cameras in the dataset, a **macro** lens (50 mm on a 35 mm **camera**) can **focus** so close that lighting remains difficult. Macro lens has a **focal** lengths from about 100 to 200 mm.

**10.STORAGE INCLUDED**:

This feature of the dataset describes about the inbuilt storage that is included in the various cameras of the dataset.

**11.WEIGHT**:

This feature of the dataset describes about the weight of the cameras(inclusive of the weight of the batteries)in the dataset.

**12.DIMENSION**:

This feature of the dataset describes about the dimensions of the cameras in the dataset.

**13.PRICE:**

This feature of the dataset describes about the prices of the cameras in the dataset.

***Preprocessing or Data Cleaning***

Datasets usually contain large volumes of data that may be stored in formats that are not easy to use. Data cleaning is the process of detecting and correcting corrupt or inaccurate records from a record set, table, or database.

For Data Cleaning we first analyzed each of the features and found out the total number of missing values. Then according to the number of missing values we have replaced numeric values with mean so that there is no bias in the dataset. We have no categorical missing values in our dataset. After handling the missing data we have fixed inconsistent capitalization. Then we fixed all the typos present in the dataset by finding and replacing the corrected typos respectively. The new dimensions of the dataset after and before cleaning are the same and are **1038 rows and 13 columns.**

Missing and erroneous data can pose a significant problem to the reliability and validity of study outcomes. If data analysis is carried out on missing and erroneous data, the results obtained may vary the real outcome by a very large or small value depending on the dataset. Such problems can be avoided through data cleaning as this helps to predict outcomes more effectively without any bias which may be present in the dataset.

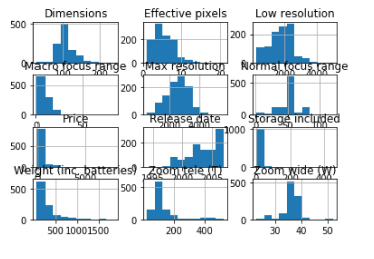
In statistics, an outlier is an observation point that is distant from other observations. The outliers can be a result of a mistake during data collection or it can be just an indication of variance in your data. To check if our dataset has outliers we have plotted a **boxplot**. If there is an outlier it will be plotted as point in the boxplot but other population will be grouped together and will be displayed as boxes. We have removed the outliers of the ‘**release date’** and **‘price’** features as they might cause a huge variation while analyzing the data. So to remove these outliers we have filtered those columns using a simple method where we find the first and third quartiles and subtract them to get the inter-quartile range. Once we obtain this, we remove all the values which do not lie in this range. After performing this method of filtration we can see that the columns do not contain any unwanted outliers.

***Exploratory Data Analysis***

Data visualization is the graphical representation of the information and data. By using the various visualization techniques we have plotted the charts and graphs which make our analysis better understandable and can see the outliers and patterns easily accessible for various analysis.

**HISTOGRAM**

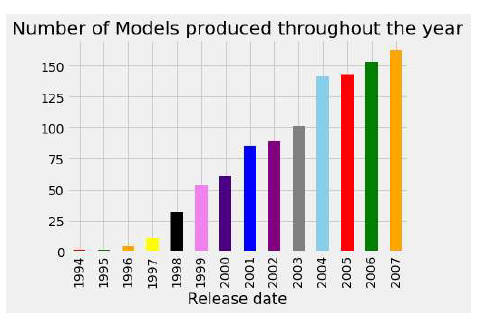
A Histogram is used to summarize discrete or continuous data. It is the most commonly used graph to show frequency distribution.



This shows the Histogram of the various features of our camera dataset.

**BAR CHART**

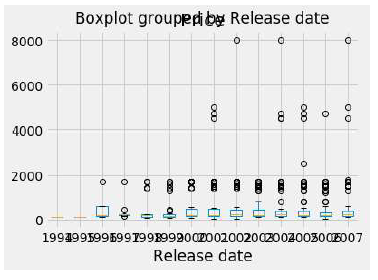
A Bar chart is a graph that represents the categorical data with rectangular bars with heights or lengths proportional to the values they represent. The bar graph shows the discrete categories of the dataset.

****

This shows the increase in the production of different models of camera in every year.

**BOX PLOT**

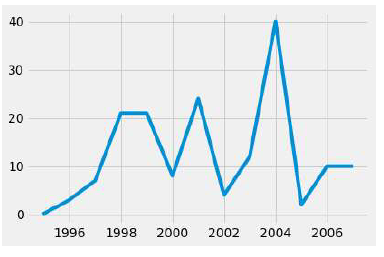
A Box plot which is also known as a whisker plot , is a type of chart often used in explanatory data analysis to visually show the distribution of numerical data through displaying the data quartiles and averages.



This shows the unwanted outliers of the dataset which is not required for our analyses

**LINE GRAPH**

Line graph is used to show the growth in the production of the camera. The line graph is used for graphical visualization of the discrete data.

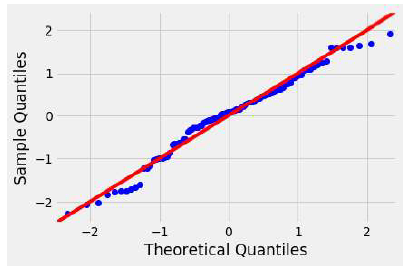


This shows the growth in production of camera in the respective year.

***Normalization and Correlation***

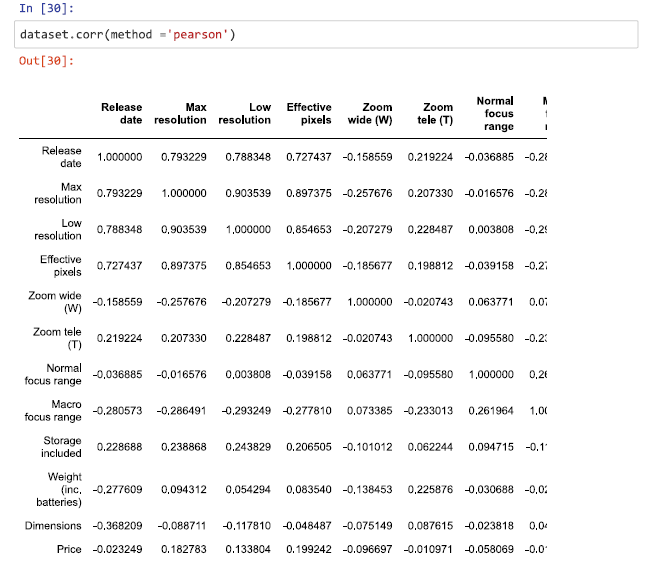
**Normalization** is a technique often applied as a part of data analyses. The goal of this is to change the values of numeric columns in the dataset to a common scale.

We have normalize all the numeric columns by making mean 0 and variance 1.



This graph shows the normalization of the mean and variance of the various features of the dataset.

**Correlation** is a powerful statistical concept that refers to a linear relationship between variables. It lies in the center of **regression** analysis techniques.



An effect score closer to 0 translates to there being no relationship.

Score closer to 1 or -1 is a positive or negative relationship respectively.

***Hypothesis testing***

Hypothesis testing is done when an assumption is tested regarding a population parameter. The methodology used depends on the nature of the data and the reason for the analysis. The null hypothesis is usually a hypothesis of equality between population parameters, in our case we have conducted three hypothesis tests as shown below:->

**HYPOTHESIS 1:**

It has been observed that the mean release date of a camera model is 2004.Discuss the validity of this statement.

**Null Hypothesis:**

The mean release date of the camera is not equal to 2004.

**EQ->H0:mean!=2004**

**Alternate Hypothesis:**

The mean release date of the camera is equal to 2004.

**EQ->H1:mean=2004**

**SAMPLE:**

SAMPLE Size(n)=5

SAMPLE VALUES={1997,1998,2000,1999,1999}

SAMPLE MEAN (X\_bar)=1998.6 SAMPLE STD=1.14

**HYPOTHESIS *2:***

It has been observed that the mean value of the dimensions of a camera model is greater than 106.Discuss the validity of this statement.

**Null Hypothesis:**

The mean dimensions value of the camera is less than or equal to 1871.

**EQ->H0:mean<=106**

**Alternate Hypothesis:**

The mean dimensions value of the camera is greater than 1781.

**EQ->H1:mean>106**

**SAMPLE:**

SAMPLE Size(n)=5

SAMPLE VALUES={95,158,0,0,128}

SAMPLE MEAN (X\_bar)=76.2 SAMPLE STD=13.04

**HYPOTHESIS 3:**

It has been observed that the mean price of a camera model is greater than 460.Discuss the validity of this statement.

**Null Hypothesis:**

The mean price of the camera is greater than or equal to 460.

**EQ->H0:mean<=460**

**Alternate Hypothesis:**

The mean price of the camera is less than 460.

**EQ->H1:mean>460**

**SAMPLE:**

SAMPLE Size(n)=5

SAMPLE VALUES={179,179,179,269,1299}

SAMPLE MEAN (X\_bar)=421 SAMPLE STD=492.36

***Results and Discussion***

To summarize we first found a dataset consisting of 13 variables (features), 1038 rows and 5% missing values. We then performed data cleaning and fixed all typos, inconsistent, capitalization, etc. To further have a better understanding of the data we are dealing with, we plotted graphs like bar chart, histogram, boxplot, line graph and scatter plot and analyzed the data to draw meaningful insights. We also filtered the outliers using the boxplot method. We then applied normalization to remove data anomalies and standardized the columns on a common scale. We found that the data is normal using graphs. We then found the correlation to find out how the data are related to each other by plotting a scatter plot. We have also performed the three hypothesis tests as mentioned above. Hence from this analyzes we can conclude that the camera production has been raising up by every year and also there is a growth in the production.