KAUNAS UNIVERSITY OF TECHNOLOGY

FACULTY OF INFORMATICS

# INTRODUCTION TO ARTIFICIAL INTELLIGENCE

# LABORATORY WORK 1

# REPORT

Student: Gurban Shukurov

Lecturer: dr. Germanas Budnikas

**Kaunas**
**2023**

1. **Select (create) a dataset to perform this and other laboratory works. Your choice must be approved by the tutor.**

   **Selected Dataset:** 1000 Cameras Dataset

   **Link:** https://www.kaggle.com/datasets/crawford/1000-cameras-dataset

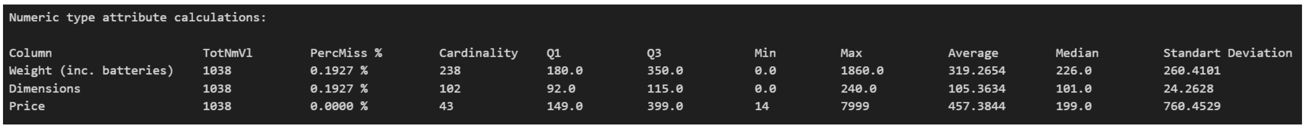
   **Context:** Based on 13 properties ~1,000 cameras were described in the dataset.

   **Format:**

   The 13 properties of each camera:

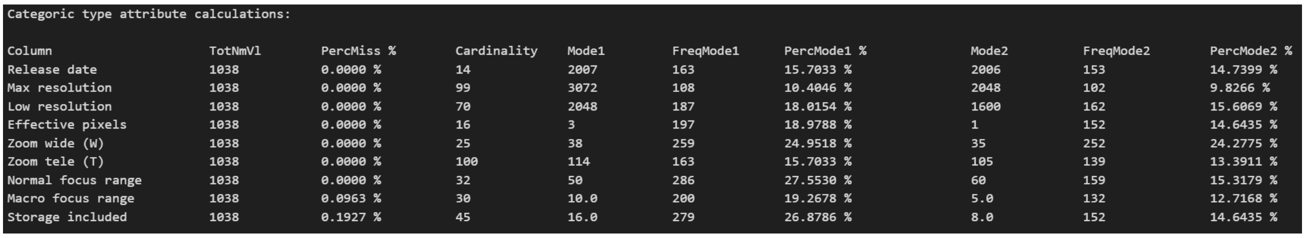
   Model
   Release date
   Max resolution
   Low resolution
   Effective pixels
   Zoom wide (W)
   Zoom tele (T)
   Normal focus range
   Macro focus range
   Storage included
   Weight (inc. batteries)
   Dimensions
   Price

2. **For each *continuous* (numeric) type attribute calculate:**
   - total number of values,
   - percentage of missing values,
   - cardinality,
   - minimum (min) and maximum (max) values,
   - 1st and 3rd quartiles,
   - average,
   - median,
   - Standard deviation.



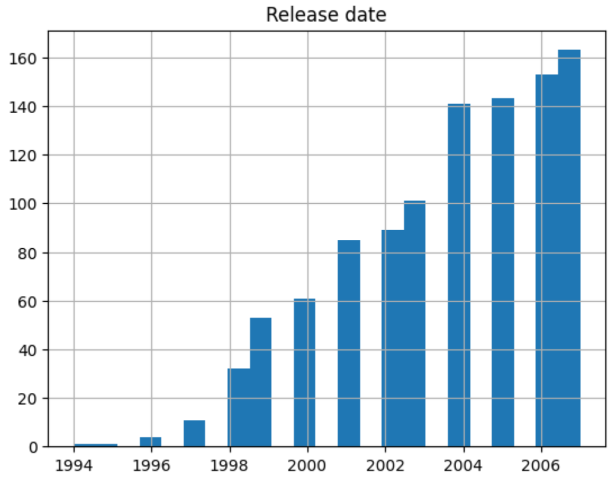
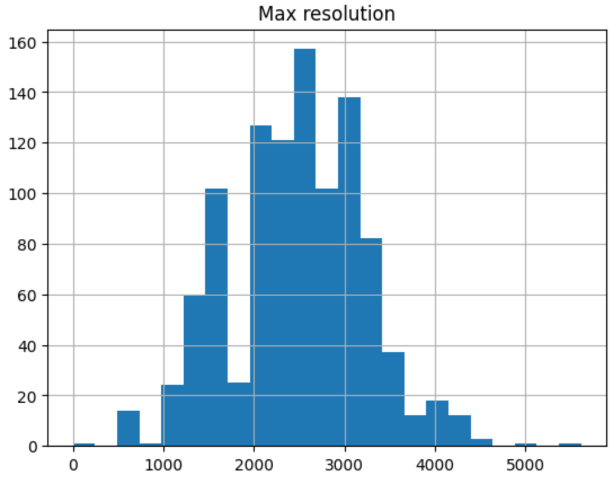
```
Numeric type attribute calculations:

Column                   TotNmVl     PercMiss %     Cardinality  Q1        Q3        Min      Max       Average    Median     Standart Deviation
Weight (inc. batteries)  1038        0.1927 %       238          180.0     350.0     0.0      1860.0    319.2654   226.0      260.4101
Dimensions               1038        0.1927 %       102          92.0      115.0     0.0      240.0     105.3634   101.0      24.2628
Price                    1038        0.0000 %       43           149.0     399.0     14       7999      457.3844   199.0      760.4529
```

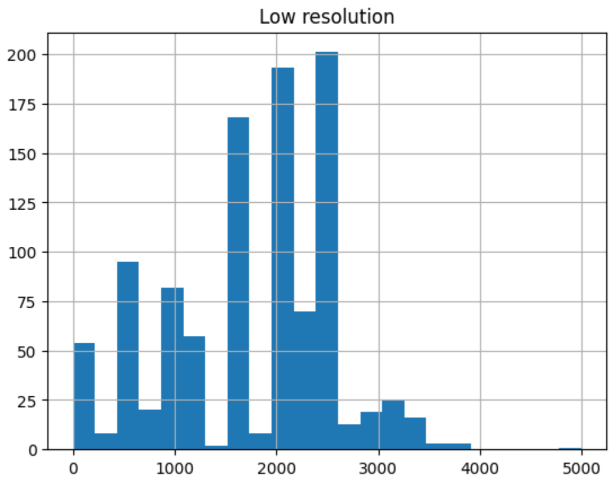
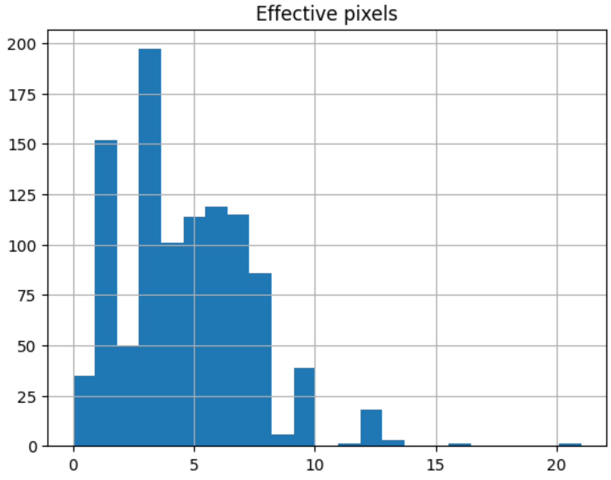
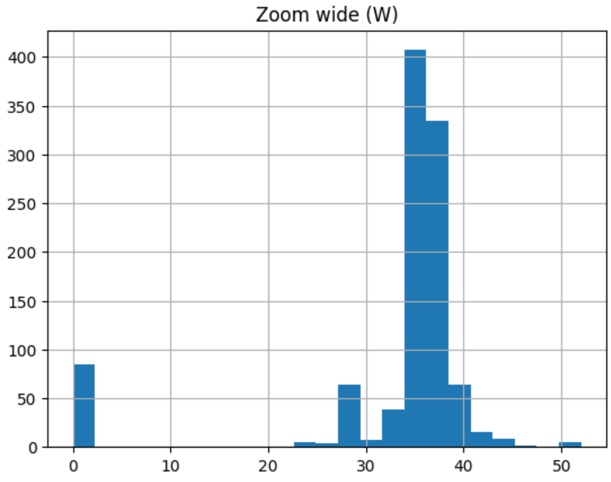
3. **For each *category* type attribute calculate:**
   - total number of values,
   - percentage of missing values,
   - cardinality,
   - mode,
   - The frequency of the mode
   - Percentage value of the mode
   - Second mode value (mode 2),
   - Frequency value for Mode 2,
   - Percentage of Mode 2.

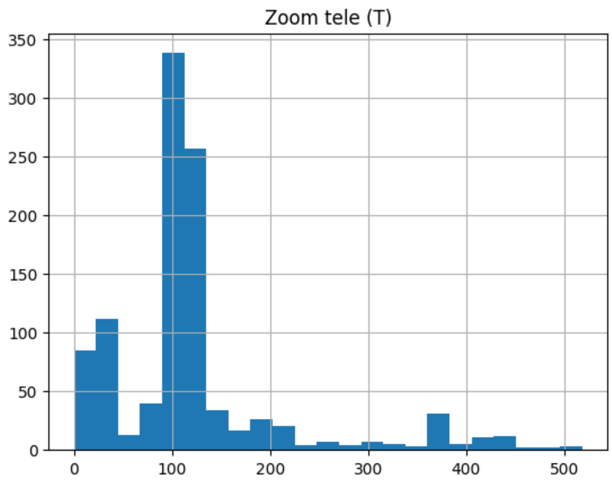
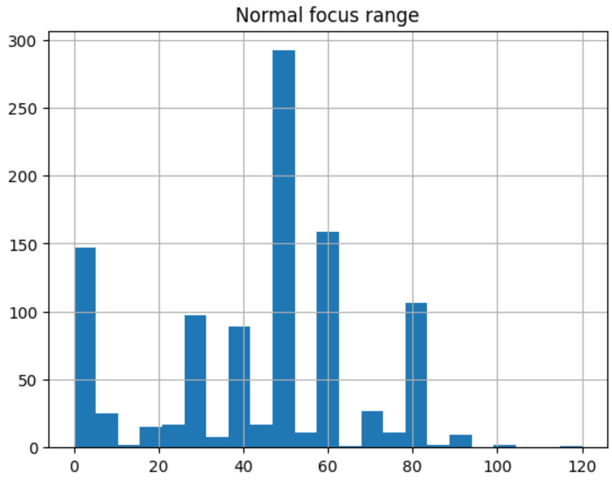
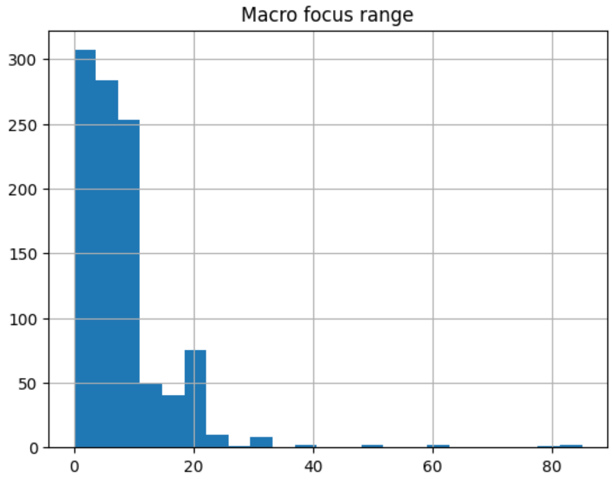


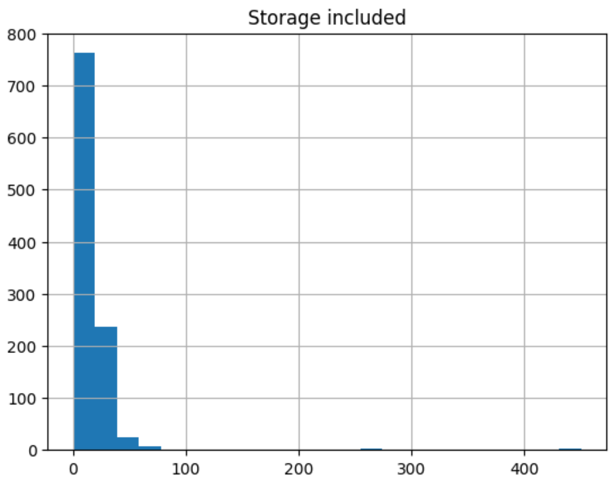
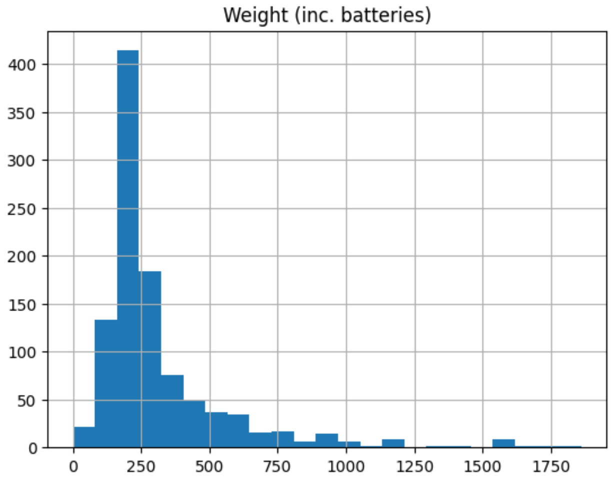
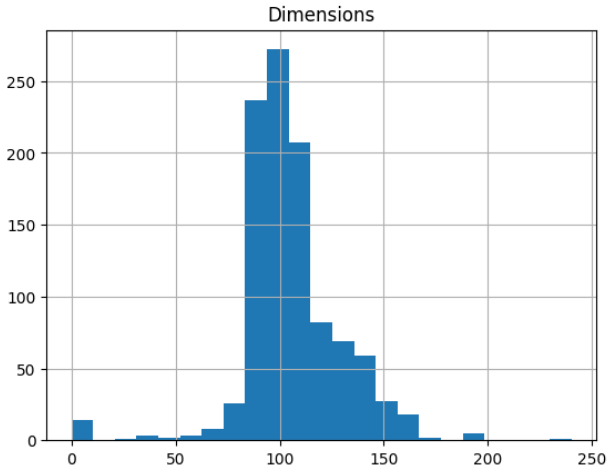
```
Categoric type attribute calculations:

Column                     TotNmVl        PercMiss %        Cardinality    Mode1         FreqMode1      PercMode1 %              Mode2          FreqMode2        PercMode2 %
Release date               1038           0.0000 %          14             2007          163            15.7033 %                2006           153              14.7399 %
Max resolution             1038           0.0000 %          99             3072          108            10.4046 %                2048           102              9.8266 %
Low resolution             1038           0.0000 %          70             2048          187            18.0154 %                1600           162              15.6069 %
Effective pixels           1038           0.0000 %          16             3             197            18.9788 %                1              152              14.6435 %
Zoom wide (W)              1038           0.0000 %          25             38            259            24.9518 %                35             252              24.2775 %
Zoom tele (T)              1038           0.0000 %          100            114           163            15.7033 %                105            139              13.3911 %
Normal focus range         1038           0.0000 %          32             50            286            27.5530 %                60             159              15.3179 %
Macro focus range          1038           0.0963 %          30             10.0          200            19.2678 %                5.0            132              12.7168 %
Storage included           1038           0.1927 %          45             16.0          279            26.8786 %                8.0            152              14.6435 %
```

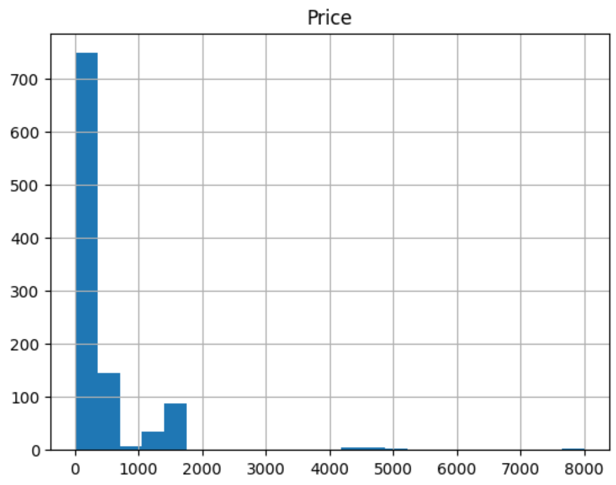
4. **Draw histograms of attributes (recommended number of histogram columns is defined by a formula: 1+3.22·$\log_e n$ , where n is sample size). Provide descriptions of the distribution (e.g., normal, exponential, etc.) and what conclusions can be drawn from it.**

| **Histogram** | **Description** |
|---|---|
|  | Exponential distribution<br>Outlier between 1994 and 1996 |
|  | Normal distribution<br>Outliers between 0 and 500 and after 4500 |

| | |
|---|---|
|  | Lognormal distribution<br>Outlier between 4500 and 5000 |
|  | Outliers after 10 |
|  | Outliers between 0 and 10, after 50 |

| | |
|---|---|
|  | Looks like normal distribution, but is not |
|  | Outliers after 100 |
|  | Normal distribution<br>Outliers after 30<br>Skewed right |

| | |
|---|---|
|  | Normal distribution.<br>Outliers between 200 and 300, after 400.<br>Skewed right |
|  | Normal distribution.<br>Outliers after 1250<br>Skewed right |
|  | Normal distribution<br>Outliers between 0 and 25, 175 and 250 |

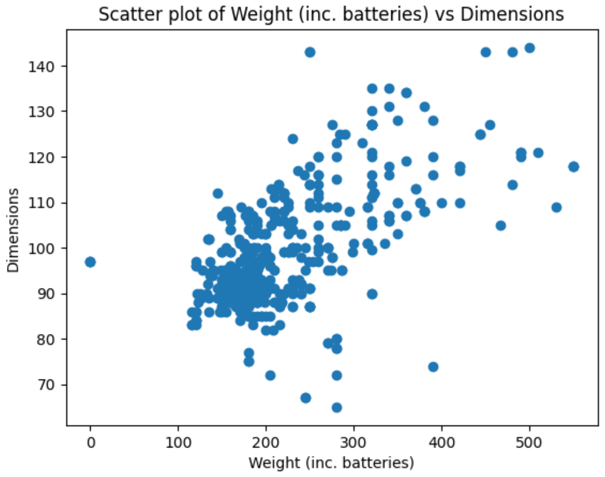
| | |
|---|---|
|  | Outliers between 4000 and 6000, 7000 and 8000 |

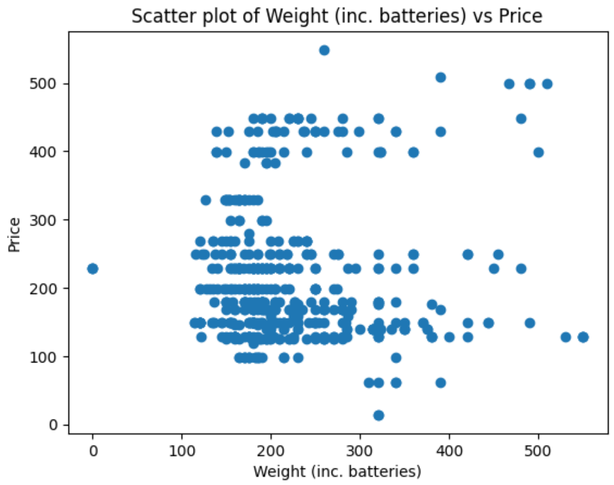
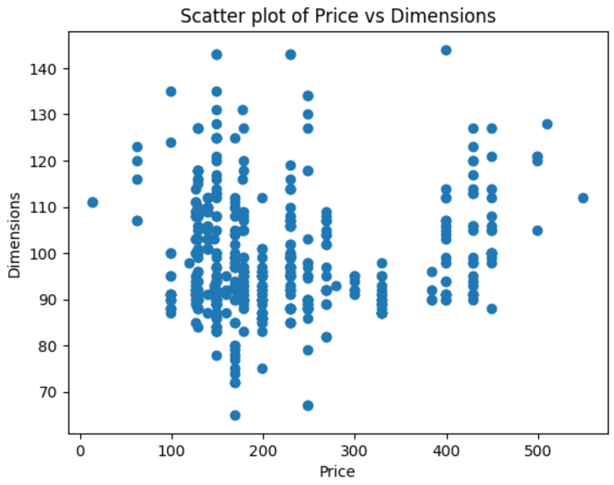
5. **Identify data quality problems: missing values, cardinality problems, outliers. Provide a plan for resolving these issues, which will be implemented programmatically (e.g. missing categorical attribute values will be included based on an attribute's mode estimate, extreme values are eliminated or adjusted).**

   There were some missing values and I filled them with the mode of the attribute, accordingly.
   There were no cardinality problems.
   In the histograms I marked the outliers, they were supposed to be removed, so I did.
   I did the following tasks without data quality problems (i.e., with corrected dataset).

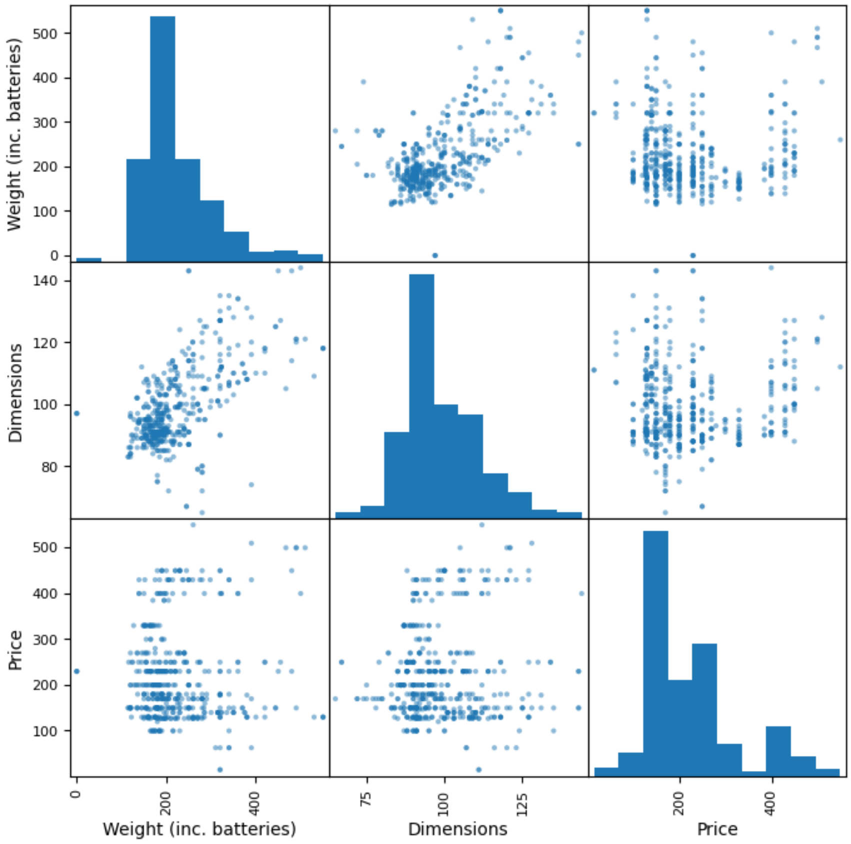
6. **Investigate relationships between attributes using visualization techniques**
   a. **For continuous type attributes: Using a scatter plot type graph, provide *several* (2-3) examples with strong linear attribute dependency (direct or inverse correlation) and *several* examples with non-correlated (weakly correlated) attributes. Comment on results.**
   b. **Provide an SPLOM diagram (Scatter Plot Matrix).**

   In my dataset I had only three numerical attributes

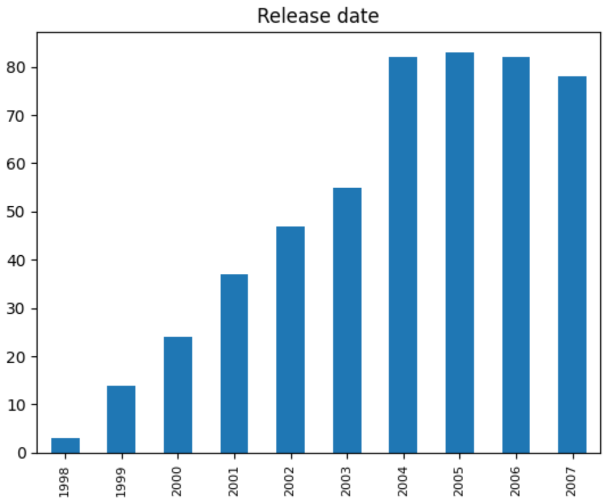
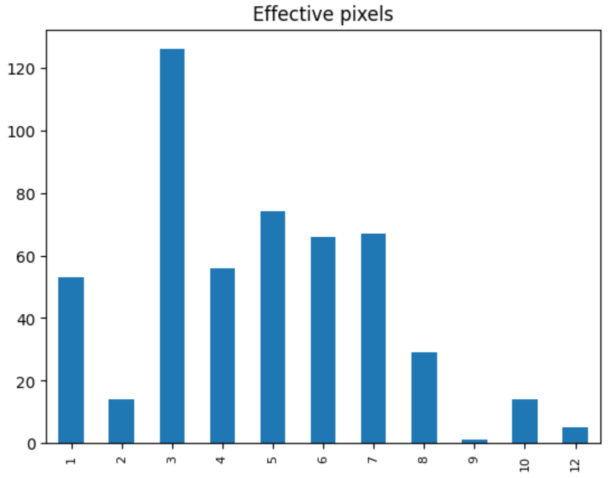
| **Scatter Plot** | **Description** |
|---|---|
|  | Linear correlation.<br>As the size of camera was increasing, weight was increasing accordingly |

| | |
|---|---|
|  | No correlation between weight and price |
|  | No correlation between price and dimensions |

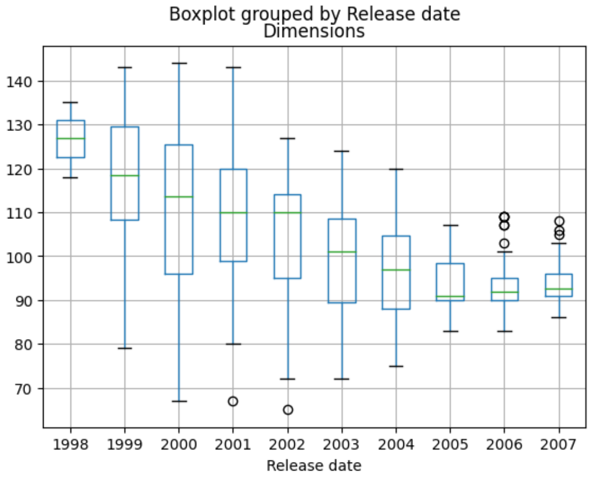
**SPLOM-Diagram:**

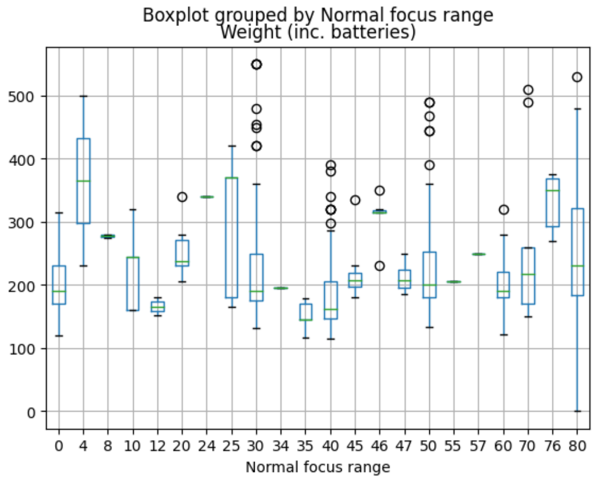


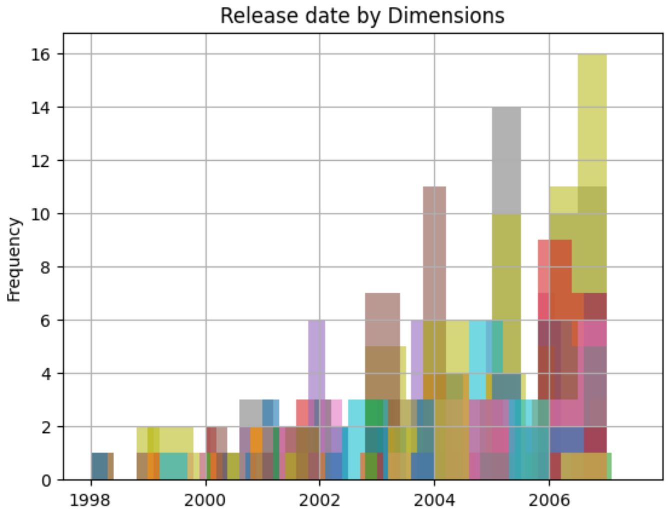
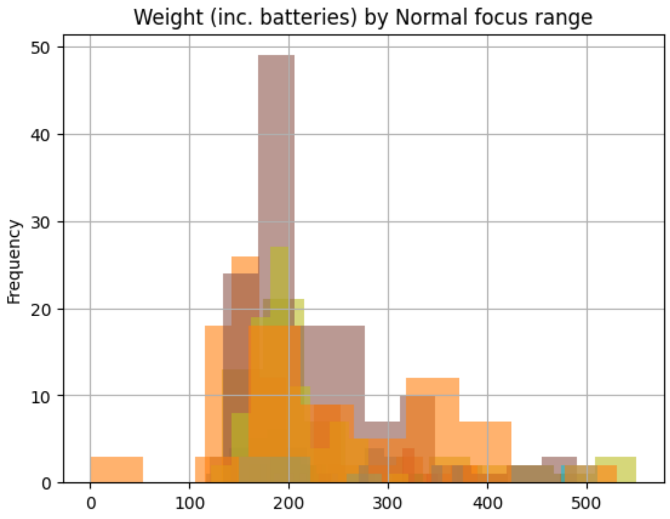
- ***For categorical attributes:*** **Using the bar plot type diagram, give some (2-3) examples of attribute frequency and comment on the results.**

| **Bar plots** | **Comment** |
|---|---|
|  | Year by year the number of manufactured cameras was increasing |
|  | Most of the cameras have only 3 effective pixels |

- **Provide some (2-3) examples of histograms and box plot diagrams depicting relationships between categorical and numeric type variables.**

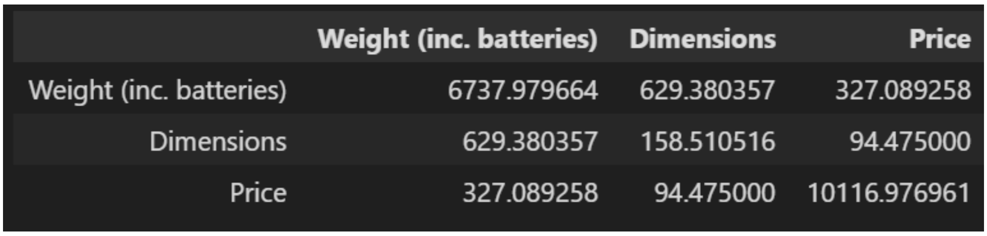
| **Boxplot** | **Description** |
|---|---|
|  | Dimensions of cameras were being reduced year by year |

| | |
|---|---|
|  | Many types of cameras had normal focus range of 80,60,50,40 that weighted 200-300 |

| Histogram | Description |
|---|---|
|  | Dimensions were decreasing after release dates |
|  | Normal focus range was increasing until the weight became 200, then it started decreasing |

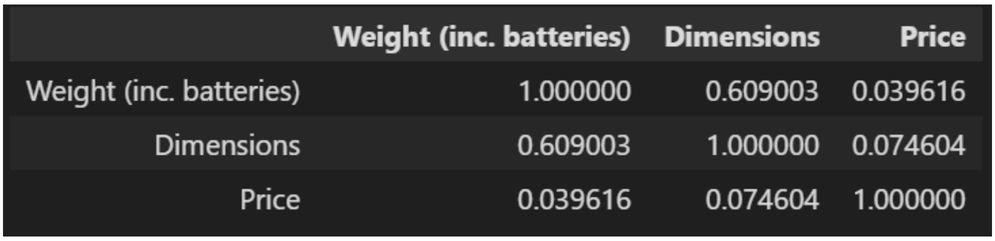
7. **Calculate the covariance and correlation values between continuous attributes and graphically represent the correlation matrix. Comments on the results.**

**Covariance: The values in datasets are very different and unique**



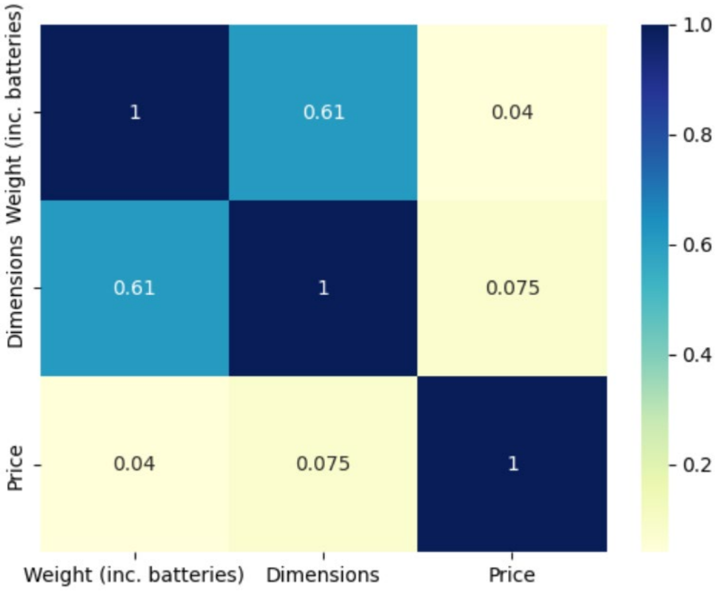
| | Weight (inc. batteries) | Dimensions | Price |
|---|---|---|---|
| Weight (inc. batteries) | 6737.979664 | 629.380357 | 327.089258 |
| Dimensions | 629.380357 | 158.510516 | 94.475000 |
| Price | 327.089258 | 94.475000 | 10116.976961 |

**Correlation: it was positive correlation, but not perfect one**



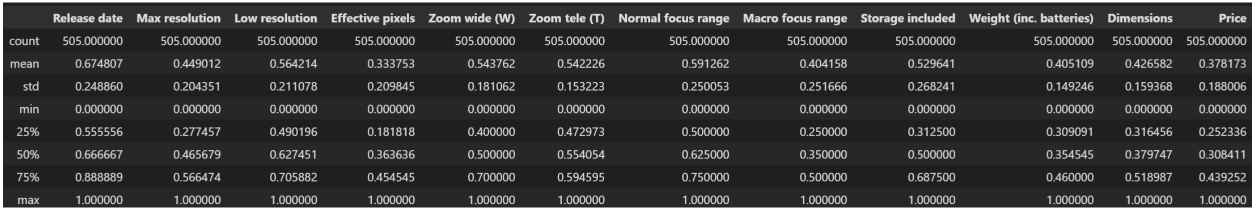
| | Weight (inc. batteries) | Dimensions | Price |
|---|---|---|---|
| Weight (inc. batteries) | 1.000000 | 0.609003 | 0.039616 |
| Dimensions | 0.609003 | 1.000000 | 0.074604 |
| Price | 0.039616 | 0.074604 | 1.000000 |

**Correlation matrix:**



### 8. Perform data normalization.

I performed data normalization in the bounds of 0 and 1:



| | Release date | Max resolution | Low resolution | Effective pixels | Zoom wide (W) | Zoom tele (T) | Normal focus range | Macro focus range | Storage included | Weight (inc. batteries) | Dimensions | Price |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 505.000000 | 505.000000 | 505.000000 | 505.000000 | 505.000000 | 505.000000 | 505.000000 | 505.000000 | 505.000000 | 505.000000 | 505.000000 | 505.000000 |
| mean | 0.674807 | 0.449012 | 0.564214 | 0.333753 | 0.543762 | 0.542226 | 0.591262 | 0.404158 | 0.529641 | 0.405109 | 0.426582 | 0.378173 |
| std | 0.248860 | 0.204351 | 0.211078 | 0.209845 | 0.181062 | 0.153223 | 0.250053 | 0.251666 | 0.268241 | 0.149246 | 0.159368 | 0.188006 |
| min | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 0.555556 | 0.277457 | 0.490196 | 0.181818 | 0.400000 | 0.472973 | 0.500000 | 0.250000 | 0.312500 | 0.309091 | 0.316456 | 0.252336 |
| 50% | 0.666667 | 0.465679 | 0.627451 | 0.363636 | 0.500000 | 0.554054 | 0.625000 | 0.350000 | 0.500000 | 0.354545 | 0.379747 | 0.308411 |
| 75% | 0.888889 | 0.566474 | 0.705882 | 0.454545 | 0.700000 | 0.594595 | 0.750000 | 0.500000 | 0.687500 | 0.460000 | 0.518987 | 0.439252 |
| max | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |

### 9. Convert categorical variables to numeric type variables.

I did not have to convert the values in my dataset because they were already numerical.