**EKONOMETRIA PROJEKT ZALICZENIOWY**

BUDOWA MODELU EKONOMETRYCZNEGO

Autorzy: Antoni Kois, Fili Mazur

Opis pracy: Badanie czynników wpływających na cenę telefonu

Spis treści

[Dane – źródło i opis 3](#_Toc197182500)

[Źródło 3](#_Toc197182501)

[Opis 3](#_Toc197182502)

[Podstawowe statystyki 3](#_Toc197182503)

[Analiza współczynników korelacji pomiędzy zmiennym 4](#_Toc197182504)

[Przygotowanie danych 6](#_Toc197182505)

[Odrzucenie pierwszych zmiennych oraz zmiana formy danych 6](#_Toc197182506)

[Zamiana danych “Yes”/”No” na zmienne zero-jedynkowe 6](#_Toc197182507)

[Usunięcię wartości odstających 7](#_Toc197182508)

# Dane – źródło i opis

## Źródło

Dane zawierające ceny laptopów oraz potencjalne zmienne mające wpływ na cenę. Zostały one pobrane ze strony kaggle.com. Podzeliliśmy je na dwa zbiory, “train” do budowy modelu oraz “test” do testowania stworzonego modelu. Dane do budowy modelu zawierają 1000 wierszy, natomiast do testowania 200+ wierszy  
  
[Link do źródła danych](https://www.kaggle.com/datasets/owm4096/laptop-prices)

## Opis

* **Company -** Marka (firma) laptopa
* **TypeName -** Typ laptopa (Notebook, Ultrabook, Gaming itp.)
* **Inches -** Przekątna ekranu (cale)
* **Ram -** Pamięć RAM (GB)
* **OS -** System operacyjny
* **Weight -** Masa laptopa (kg)
* **Price\_euros -** Cena laptopa (euro) – zmienna docelowa
* **Screen -** Rozdzielczość/definicja ekranu (Standard, Full HD, 4K Ultra HD, Quad HD+)
* **ScreenW -** Szerokość ekranu w pikselach
* **ScreenH -** Wysokość ekranu w pikselach
* **Touchscreen -** Czy ekran jest dotykowy (0 = nie, 1 = tak)
* **IPSpanel -** Czy ekran ma matrycę IPS (0 = nie, 1 = tak)
* **RetinaDisplay -** Czy ekran to Retina/ekran o bardzo wysokiej gęstości pikseli
* **CPU\_company -** Producent procesora (np. Intel, AMD)
* **CPU\_freq -** Taktowanie procesora (Hz)
* **PrimaryStorage -** Pojemność głównego dysku (GB)
* **SecondaryStorage -** Pojemność dodatkowego dysku (GB, jeśli występuje)
* **PrimaryStorageType -** Typ głównego dysku (HDD, SSD, Flash Storage, Hybrid)
* **SecondaryStorageType -** Typ dodatkowego dysku (HDD, SSD, Hybrid, None)
* **GPU\_company -** Producent karty graficznej

## Podstawowe statystyki

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Zmienna | Średnia | Odchylenie standardowe | Kurtoza | Skośność | Współczynnik zmienności |
| Inches | 15.0498 | 1.413668 | -0.19787 | -0.396643 | 0.093933 |
| Ram | 8.512 | 4.973791 | 6.300672 | 2.074162 | 0.584327 |
| Weight | 2.037788 | 0.662631 | 2.516759 | 1.174558 | 0.325172 |
| Price\_euros | 1125.70607 | 705.380659 | 5.046084 | 1.620813 | 0.626612 |
| ScreenW | 1908.702 | 495.832344 | 6.758238 | 2.277741 | 0.259775 |
| ScreenH | 1079.172 | 285.957952 | 5.93948 | 2.17667 | 0.264979 |
| CPU\_freq | 2.2868 | 0.525055 | -0.434995 | -0.723646 | 0.229603 |
| PrimaryStorage | 444.428 | 372.158373 | 3.155327 | 1.624314 | 0.837387 |
| SecondaryStorage | 179.956 | 427.102459 | 4.757773 | 2.311719 | 2.373372 |

## Analiza współczynników korelacji pomiędzy zmiennym

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Zmienna | Inches | Ram | Weight | Price\_euros | ScreenW | ScreenH | CPU\_freq | PrimaryStorage | SecondaryStorage |
| Inches | 1.00000 | 0.23295 | 0.82392 | 0.07079 | -0.04676 | -0.07043 | 0.29613 | 0.26112 | 0.38299 |
| Ram | 0.23295 | 1.00000 | 0.38976 | 0.76340 | 0.45563 | 0.44695 | 0.37303 | 0.01037 | 0.39085 |
| Weight | 0.82392 | 0.38976 | 1.00000 | 0.21953 | -0.00383 | -0.02567 | 0.31324 | 0.17174 | 0.47008 |
| Price\_euros | 0.07079 | 0.76340 | 0.21953 | 1.00000 | 0.57717 | 0.57513 | 0.43245 | -0.10604 | 0.28036 |
| ScreenW | -0.04676 | 0.45563 | -0.00383 | 0.57717 | 1.00000 | 0.99355 | 0.18731 | -0.05033 | 0.15056 |
| ScreenH | -0.07043 | 0.44695 | -0.02567 | 0.57513 | 0.99355 | 1.00000 | 0.17671 | -0.05732 | 0.13869 |
| CPU\_freq | 0.29613 | 0.37303 | 0.31324 | 0.43245 | 0.18731 | 0.17671 | 1.00000 | 0.05396 | 0.20782 |
| PrimaryStorage | 0.26112 | 0.01037 | 0.17174 | -0.10604 | -0.05033 | -0.05732 | 0.05396 | 1.00000 | -0.23088 |
| SecondaryStorage | 0.38299 | 0.39085 | 0.47008 | 0.28036 | 0.15056 | 0.13869 | 0.20782 | -0.23088 | 1.00000 |

A chart with different colored squares

Description automatically generated

# Przygotowanie danych

## Odrzucenie pierwszych zmiennych oraz zmiana formy danych

Na wstępie nastepujące zmienne zostają odrzucone bądź zostają do dalszej analizy:

- Company - zostawiamy top 6 kategorii co obejmuje 90.30% wierszy (19 kategorii)

- Product - zbyt dużo kategorii (618) i zmienna jest mniej istotna

- TypeName - mniej istotna zmienna

- Inches - pozozstawiamy

- Ram - pozozstawiamy

- OS - zostawiamy top 5 kategorii co obejmuje 97.60% wierszy (9 kategorii)

- Weight - pozozstawiamy

- Price\_euros - pozozstawiamy

- Screen - 4 kategorie (wszystkie)

- ScreenW - pozozstawiamy

- ScreenH - pozozstawiamy

- Touchscreen – pozozstawiamy, zmienna 0/1

- IPSpanel – pozozstawiamy, zmienna 0/1

- RetinaDisplay – pozozstawiamy, zmienna 0/1

- CPU\_company - 3 kategorie (wszystkie)

- CPU\_freq - pozozstawiamy

- CPU\_model - zbyt dużo kategorii (93) i mniej istotne

- PrimaryStorage - pozozstawiamy

- SecondaryStorage - pozozstawiamy

- PrimaryStorageType - 4 kategorie (wszytskie)

- SecondaryStorageType - mała ilośc danych posiada uzupełnioną tą wartość, pomijamy

- GPU\_company - 3 kategorie (wszystkie)

- GPU\_model - zbyt dużo kategorii i mniej istotne (110)

## Zamiana danych “Yes”/”No” na zmienne zero-jedynkowe

Kolumny:

* **Touchscreen**
* **IPSpanel**
* **RetinaDisplay**

Zawierają dane w formacie napisów “Yes”/”No” dlatego, żeby były użyteczne dla naszego modelu zamieniamy je na zmienne zero-jedynkowe, czyli ich wartości będą 0 albo 1. 0 w przypadku gdy zmienna zwierała “No” oraz 1, gdy zawierała “Yes”.

## Usunięcię wartości odstających

Do ustalenia wartości odstających zostały użyte kwartyle i reguła 1,5\*IQR

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Zmienna | odstających | % | dolna | górna |
| ScreenH | 331 | 33.1 | 1080.0 | 1080.0 |
| ScreenW | 329 | 32.9 | 1920.0 | 1920.0 |
| PrimaryStorage | 203 | 20.3 | -128.0 | 896.0 |
| Ram | 171 | 17.1 | -2.0 | 14.0 |
| SecondaryStorage | 164 | 16.4 | 0.0 | 0.0 |
| Weight | 33 | 3.3 | 0.3 | 3.5 |
| Inches | 30 | 3.0 | 11.6 | 18.0 |
| Price\_euros | 29 | 2.9 | -713.875 | 2787.125 |
| CPU\_freq | 0 | 0.0 | 0.45 | 4.05 |

Z powodu iż zmienne **ScreenH**, **ScreenW**, **PrimaryStorage**, **Ram**, **SecondaryStorage** zawierają duży odsetek wartości odstjących, usuwanie ich podległo wątpiwości. Głębsza analiza danych i informacji doprowadziła do stwierdzenia o nie usuwaniu wartości odstających dla tych kolumn. Stanowsko takie zapadło gdyż np. jak dla ScreenH 80 % laptopów ma dokładnie 1080 px wysokości. Każda inna wartość (768, 1440, 1600, …) jest traktowana jako “odstająca”, a można też patrzeć na tą zmienną jako kategoryczną.

To samo zjawisko można zauważyć dla reszty wymienionych zmiennych.

Odstające wartości zostały usunięte tylko dla **Weight** i **Inches**.

W dalszej części budowy modelu być może inne pojedyncze zmiennena zostaną poprawione o usunięcie odstających wartości.

Usuwając odstające wartości dla **Weight** i **Inches** usunięto 62 obserwacji (6.20 %).