



University of Tehran

School of Electrical and Computer Engineering



Data Analysis

Final Project

Navid Adib

810100283

فهرست

1	چکیده
2	بخش ۲ جمع آوری دیتا
2	مقدمه
2	جمع آوری اطلاعات یک محصول
2	لیست محصولات صفحه search
4	بخش سوم: پیش پردازش و تمیزسازی دیتا
7	بخش چهارم: مصورسازی و تحلیل EDA
10	بخش ششم: روش های طبقه بندی
10	طبقه بند logistic regression
10	طبقه بند SVM
11	طبقه بند Decision Tree
11	طبقه بند KNN
11	طبقه بند Random forest regression
12	طبقه بند perceptron
12	طبقه بند mlpclassifier
12	جمع بندی

هدف این پروژه کشف روابطی بین پارامترهای یک mouse و قیمت آن می باشد. لذا ابتدا دیتا را با نوشتن یک crawler از سایت amazon جمع آوری کردیم و آن را ابتدا تمیز و پیش پردازش کردیم و با رسم نمودار به بررسی روابط بین مشخصات مهم یک mouse پرداختیم و ویژگی های مهم را استخراج کردیم و سپس از مدل ها معروف یادگیری ماشین برای پیش بینی قیمت این محصول استفاده کردیم.

بخش ۲ جمع آوری دیتا

مقدمه:

در این بخش از سایت آمازون اطلاعات مربوط به محصولات mouse را جمع آوری کردیم.

جمع آوری اطلاعات یک محصول:

یک صفحه محصول در سایت Amazon به شکل زیر می باشد.

Roll over image to zoom in

Technical Details	
Brand	Logitech
Manufacturer	Logitech Asia Pacific Ltd, Unit 1003, Tower 1, Cheung Sha Wan Plaza, Cheung Sha Wan Road, Kowloon, HK
Model	910-001439
Model Name	B100
Model Year	2010
Product Dimensions	4.19 x 13.21 x 8.71 cm; 108.86 Grams
Item model number	910-001439
Computer Memory Type	DDR3 SDRAM
Operating System	Windows 10, 11 or later, Linux, Chrome OS, macOS 10.5 or later
Hardware Platform	usb
Hardware Interface	USB 2.0
Compatible Devices	Laptop, Personal Computer
Special Features	optical
Mounting Hardware	Logitech B100 Optical USB Mouse
Compiler	COPPER

Technical Details

Brand	Logitech
Manufacturer	Logitech Asia Pacific Ltd, Unit 1003, Tower 1, Cheung Sha Wan Plaza, Cheung Sha Wan Road, Kowloon, HK
Model	910-001439
Model Name	B100
Model Year	2010
Product Dimensions	4.19 x 13.21 x 8.71 cm; 108.86 Grams
Item model number	910-001439
Computer Memory Type	DDR3 SDRAM
Operating System	Windows 10, 11 or later, Linux, Chrome OS, macOS 10.5 or later
Hardware Platform	usb
Hardware Interface	USB 2.0
Compatible Devices	Laptop, Personal Computer
Special Features	optical
Mounting Hardware	Logitech B100 Optical USB Mouse
Compiler	COPPER

Additional Information

ASIN	B003L62T7W
Customer Reviews	★★★★☆ 31,711 ratings 4.3 out of 5 stars
Best Sellers Rank	#459 in Electronics (See Top 100 in Electronics) #13 in Mice
Date First Available	5 July 2012
Packer	Logitech Asia Pacific Ltd, Unit no. 1003, Tower 1, Cheung Sha Wan Plaza, Cheung Sha Wan Road, Kowloon, Hong Kong.
Importer	Savex Technologies Private Limited, S1, Gala No:1-2, 101-102, Perana Industrial Complex, Anjur Phata- Mankholi Rd Village, Bhiwandi District, Thane, 421302, Maharashtra
Item Dimensions LxWxH	42 x 132 x 87 Millimeters

Feedback

Would you like to [tell us about a lower price?](#)

لذا تابعی به نام `oneProductExtraction.py` نوشتیم که این اطلاعات را به کمک `bs4` استخراج کند.

لیست محصولات صفحه `search`:

در این بخش لیست `link` های موجود در صفحه `search` را پیدا کرده و به تابع قسمت قبل داده تا وارد تک تک این صفحات شود و از آن ها اطلاعات را استخراج کند.

Brand

- ☐ ZEBRONICS
- ☐ Logitech
- ☐ HP
- ☐ Portronics
- ☐ Lenovo
- ☐ Dell
- ☐ Razer
- [See more](#)

Price

Under ₹1,000
 ₹1,000 - ₹5,000
 ₹5,000 - ₹10,000
 ₹10,000 - ₹20,000
 Over ₹20,000

Min Max Go

Deals & Discounts
 Today's Deals

Computers & Accessories Brands

- ☐ Made for Amazon
- ☐ Top Brands


Pay On Delivery

- ☐ Eligible for Pay On Delivery


Input Mouse Hand Orientation

- ☐ Ambidextrous
- ☐ Left


RESULTS



Sponsored ⓘ
Logitech B100 Wired USB Mouse, 3 yr Warranty, 800 DPI Optical Tracking, Ambidextrous PC/Mac/Laptop - Black
 ★★★★★ ~ 31,711
₹269 ₹375 (28% off)
 Get it by **Today, January 23**
 FREE Delivery by Amazon



Sponsored ⓘ
Lenovo 400 Wireless Mouse, 1200DPI Optical Sensor, 2.4GHz Wireless Nano USB, 3-Button (Left,Right,Scroll) Upto 8M Left/Right & 100K Scroll clicks & 1yr Battery,...
 ★★★★★ ~ 6,407
₹620 ₹1,390 (55% off)
 ✓prime Get it by **Today, January 23**
 FREE Delivery by Amazon
 Bundles available



Amazon's Choice
Dell MS116 1000Dpi USB Wired Optical Mouse, Led Tracking, Scrolling Wheel, Plug and Play.
 ★★★★★ ~ 33,420
₹3,420 ₹3,999 (14% off)
 Get it by **Thursday, January 26**
 FREE Delivery by Amazon

و در نهایت بر روی button (next) کلیک کند تا لیست جدید محصولات ظاهر شود.

RELATED SEARCHES

Q mouse wireless	Q mouse pad	Q keyboard
Q gaming mouse	Q logitech mouse	

< Previous 1 2 3 ... 20 Next >

در پیاده سازی به علت نوسانات اینترنت و block شدن ip توسط خود سایت amazon از try except استفاده کردیم تا با خطا مواجه نشویم و اطلاعات را تا جایی که بدست آمده ذخیره کنیم. همچنین یکی از مشکلات اصلی این بود که ویژگی های موجود برای هر محصول در بعضی های موجود نبود و نیاز است آن ها را در نظر گرفت. در نهایت اطلاعات را با هم concatenate کرده و به all-mouse-data.csv رسیدیم.

بخش سوم: پیش پردازش و تمیزسازی دیتا

ابتدا dataframe را لود کرده و آن را باهم می بینیم.

	title	brand	colour	connectivityTechnology	specialFeature	movementDetectionTechnology	numberOfItems	price	weight	country
690	HP Bluetooth Mouse 250/4.2 Bluetooth connectiv...	HP	Black	NaN	Wireless, Bluetooth, Ergonomic, Optical	Optical	NaN	759.	70 g	NaN
691	HP 150 Wireless USB Mouse with Ergonomic and a...	HP	Black	USB	Wireless, Ergonomic Design	Optical	NaN	449.	50 g	China

ستون ها عبارتند از: title, color, connectivity technology, special feature, movement detection technology, number of items, weight, country در title و special feature اطلاعاتی نهفته است که می توان از آن ها ستون های دیگری ایجاد کرد و یا بعضی از مقادیر nan را مقدار دهی کرد. وضعیت فعلی این data frame به شکل زیر است.

#	Column	Non-Null Count	Dtype
0	title	700 non-null	object
1	brand	666 non-null	object
2	colour	652 non-null	object
3	connectivityTechnology	192 non-null	object
4	specialFeature	596 non-null	object
5	movementDetectionTechnology	607 non-null	object
6	numberOfItems	37 non-null	float64
7	price	681 non-null	object
8	weight	665 non-null	object
9	country	650 non-null	object

dtypes: float64(1), object(9)
memory usage: 54.8+ KB

ستون number of item مقدار null زیادی دارد و با دستور value_counts می بینیم که همه آن ها یک عدد است و لذا این ستون را drop می کنیم.

```
df['numberOfItems'].value_counts()
✓ 0.1s
1.0    37
Name: numberOfItems, dtype: int64
```

همچنین نمونه های تکراری را با دستور drop_duplicate حذف کردیم. و متوجه شدیم تعداد زیادی از آن ها تکراری بوده. همچنین چون لیبل اصلی قیمت است باید null نباشد و آن هایی که null هستند را حذف کردیم. همچنین تمام ستون ها را به حالت lowercase می بریم. در آخر تابعی نوشتیم که از ستون special feature و title اطلاعات دیگری چون portable بودن، ergonomic بودن و ... را استخراج کند. خلاصه این تابع را در زیر ملاحظه می کنید.

```

tractFeature(data):
    = data.copy()
    ['wireless']=False
    ['usb']=False
    ['bluetooth']=False
    ['ergonomic']=False
    ['portable']=False
    ['soundless']=False
    ['led_lights']=False
    ['rechargeable']=False
    ['lightweight']=False
    ['programmable_button']=False

r ind in df.index:
    feature = str(df['specialFeature'][ind])
    title = str(df['title'][ind])
    connTech = str(df['connectivityTechnology'][ind])
    # -----
    df['wireless'][ind] = feature.find('wireless')>-1 or title.find('wireless')>-1 or connTech.find('wireless')>-1
    df['usb'][ind] = feature.find('usb')>-1 or title.find('usb')>-1 or connTech.find('usb')>-1
    df['bluetooth'][ind] = feature.find('bluetooth')>-1 or title.find('bluetooth')>-1 or connTech.find('bluetooth')>-1
    df['ergonomic'][ind] = feature.find('ergonomic')>-1 or title.find('ergonomic')>-1
    df['portable'][ind] = feature.find('portable')>-1 or title.find('portable')>-1
    df['soundless'][ind] = feature.find('soundless')>-1 or title.find('soundless')>-1
    df['led_lights'][ind] = feature.find('led lights')>-1 or title.find('led lights')>-1
    df['rechargeable'][ind] = feature.find('rechargeable')>-1 or title.find('rechargeable')>-1
    df['lightweight'][ind] = feature.find('lightweight')>-1 or title.find('lightweight')>-1
    df['programmable_button'][ind] = feature.find('programmable buttons')>-1 or title.find('programmable buttons')>-1

```

```

# ----- movementDetectionTechnology -----
if feature.find('optical')>-1 or title.find('optical')>-1:
    df['movementDetectionTechnology'][ind]='optical'

if feature.find('laser')>-1 or title.find('laser')>-1:
    df['movementDetectionTechnology'][ind]='laser'

# df = df[df['connectivityTechnology']!='wired']
# df = df[df['connectivityTechnology']!='wireless']
df = df.drop(['specialFeature', 'connectivityTechnology', 'colour', 'country'], axis=1).reset_index()
return(df)

df = extractFeature(df)

```

در مرحله بعدی type متغیر ها را تغییر داده و برای کاهش حجم اکثرا object ها را به category تبدیل کردیم. ستون price به علت داشتن کارکتر '، نیاز به پردازش اختصاص داشت و ستون weight به شکل زیر بود.

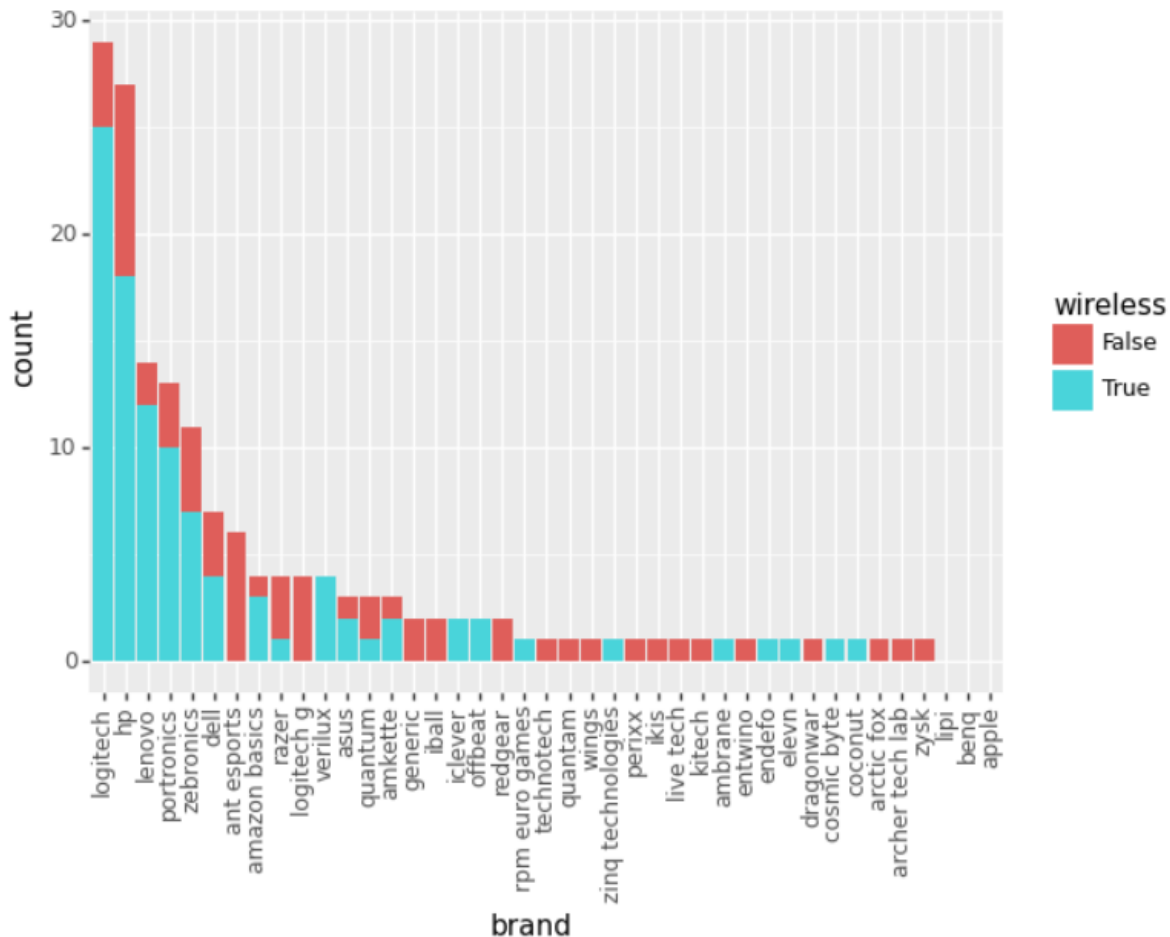
```
df['weight'].unique()
✓ 0.1s
array(['109 g ', '90 g ', '200 g ', '60.9 g ', '900 g ', '129 g ',
      '120 g ', '50 g ', '125 g ', '91 g ', '150 g ', '53 g ', '74 g ',
      '60 g ', '115 g ', '80 g ', '175 g ', '89 g ', '86 g ', '170 g ',
      '153 g ', '144 g ', '84 g ', '54 g ', '100 g ', '75 g ', '190 g ',
      '69 g ', '70 g ', '83.9 g ', '54.4 g ', '61 g ', '130 g ',
      '140 g ', '73 g ', '59.9 g ', '110 g ', nan, '77 g ', '82 g ',
      '78 g ', '96 g ', '160 g ', '134 g ', '122 g ', '55 g ', '111 g ',
      '145 g ', '68 g ', '340 g ', '105 g ', '30 g ', '65 g ', '66 g ',
      '136 g ', '94 g ', '68 Grams ', '114 g ', '85 g ', '57 g ',
      '380 g ', '27 g ', '300 g ', '83.8 g ', '83 g ', '90.7 g ',
      '1 kg 500 g ', '20 g ', '400 g ', '88 g ', '113 g ', '101 g ',
      '40 g ', '141 g '], dtype=object)
```

لذا نیاز بود g, Grams, kg را به عدد تبدیل کنیم. برای این منظور تابع g_kg_to_number را نوشتیم و به کمک متد apply و lambda function آن را اجرا کردیم و خروجی را در شکل زیر ملاحظه می کنید.

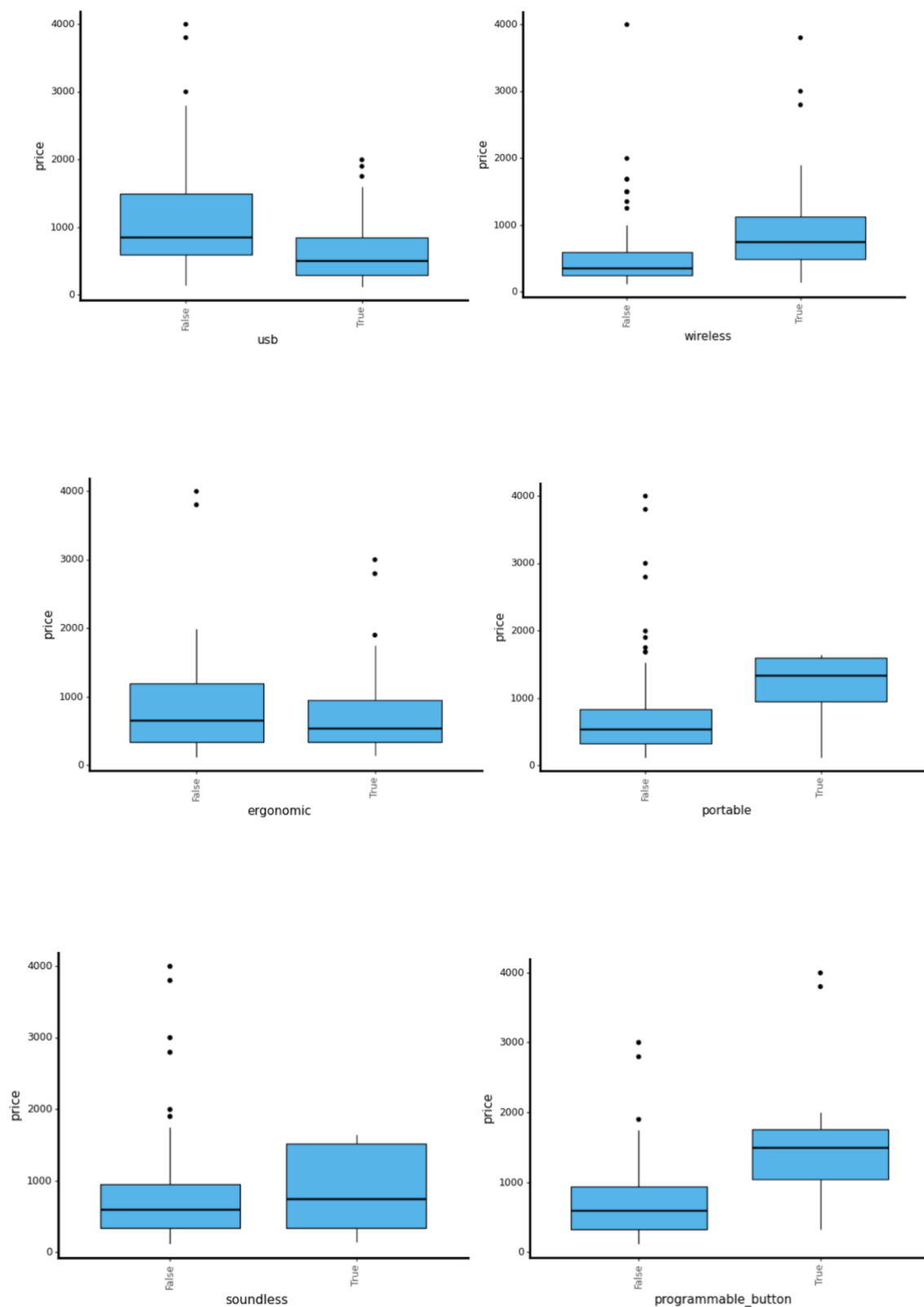
```
df['weight'].unique()
✓ 0.1s Python
array([ 109. ,  90. ,  200. ,  60.9,  900. ,  129. ,  120. ,  50. ,
        125. ,  91. ,  150. ,  53. ,  74. ,  60. ,  115. ,  80. ,
        175. ,  89. ,  86. ,  170. ,  153. ,  144. ,  84. ,  54. ,
        100. ,  75. ,  190. ,  69. ,  70. ,  83.9,  54.4,  61. ,
        130. ,  140. ,  73. ,  59.9,  110. ,  nan,  77. ,  82. ,
        78. ,  96. ,  160. ,  134. ,  122. ,  55. ,  111. ,  145. ,
        68. ,  340. ,  105. ,  30. ,  65. ,  66. ,  136. ,  94. ,
        114. ,  85. ,  57. ,  380. ,  27. ,  300. ,  83.8,  83. ,
        90.7, 1500. ,  20. ,  400. ,  88. ,  113. ,  101. ,  40. ,
        141. ])
```


بخش چهارم: مصورسازی و تحلیل EDA:

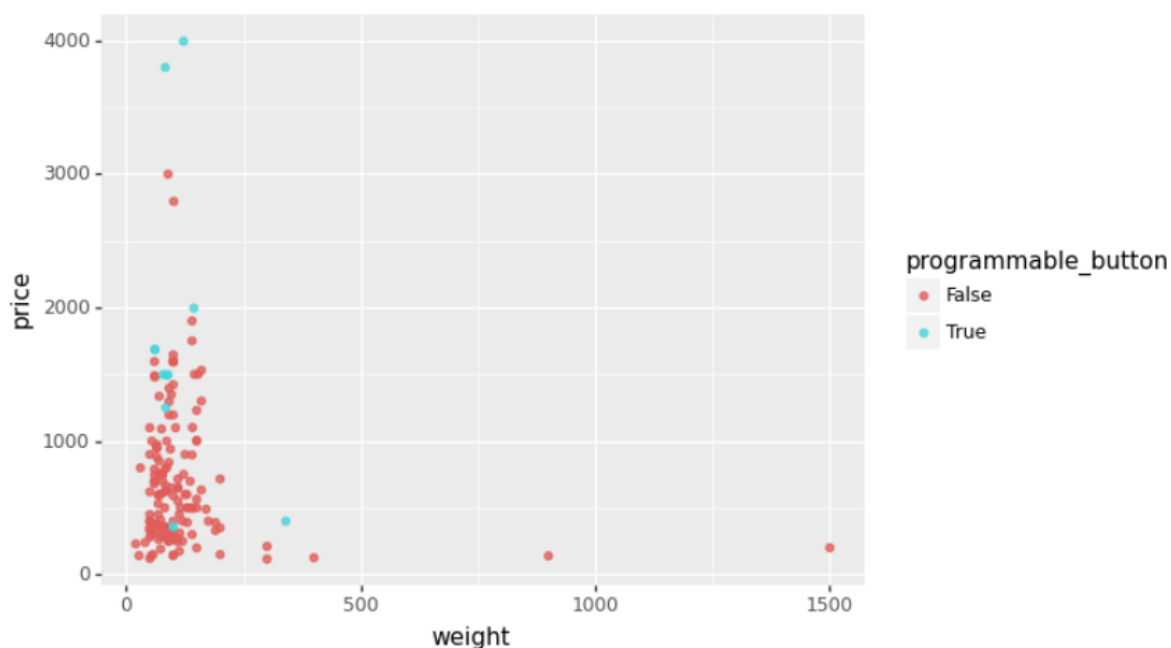
ابتدا داده ها را از نظر brand و نوع ارتباط آن ها به شکل زیر نمایش دادیم.



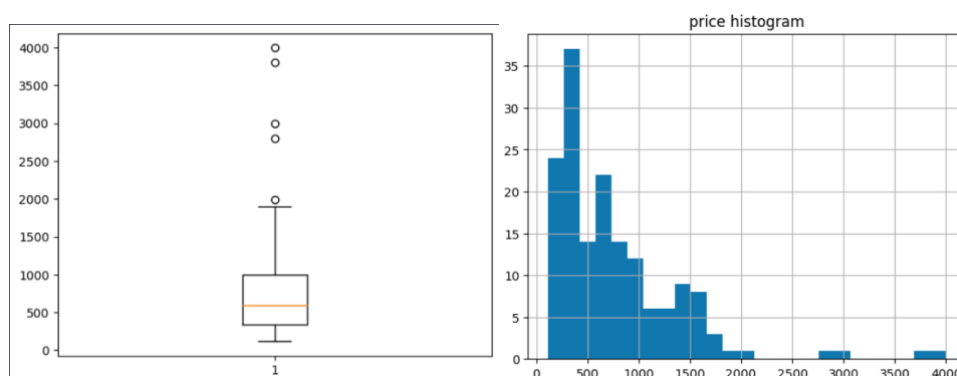
نتایج نشان میدهد brand های 'logitech'، 'hp'، 'lenovo' ... بیشترین سهم از بازار را داشتند. و در این میان نوع ارتباط از طریق wireless سهم بیشتری را در اکثر برند ها دارد. از این نوع نمایش استفاده کردیم زیرا brand یک دیتا categorical می باشد و در چنین شرایط bar plot گزینه خوبی برای نمایش است. از نمایش boxplot برای یافتن اثر انواع ویژگی ها بر روی قیمت نمودار زیر را رسم کردیم.



نتایج نشان می دهد داشتن یک رابطه همبستگی مثبت بین قیمت و متغیر های wireless, portable, programmable, soundless است. و رابطه همبستگی منفی با داشتن usb وجود دارد. در ادامه scatter plot های numerical وزن و قیمت را در دو دسته programmable و non programmable رسم کردیم.



نتایج نشان میدهد عمده وزن کمتر از ۵۰۰ گرم است و داشتن خاصیت programable واقعا در افزایش قیمت تاثیر گذار است. در انتها توزیع قیمت را به تنهایی اراعه می دهیم.



طبق صورت مسئله پروژه نیاز است تا لیبیل پیوسته را گسسته نمود لذا با توجه به median و histogram قیمت محصولات ترشلد را در ۶۰۰ قرار دادیم. و به دو لیبیل low و high رسیدیم.

بخش ششم: روش های طبقه بندی

طبقه بند logistic regression:

در این بخش از k-fold cross validation برای اعتبار سنجی مدل با $k=8$ استفاده کردیم. همچنین از standard scaler و logistic regression به صورت متوالی و در یک pipeline استفاده کردیم. همچنین hyperparameter c در مدل logistic regression را به ازای مقادیر مختلف تست کرده و بهترین را قرار دادیم. نتایج را در زیر ملاحظه میکنید.

```
best accuracy of logisticregression with 8-hold cross validation is:
79.52380952380953 %
hyper parameter: c = 0.1
-----
fit_time 0.028614461421966553
score_time 0.026972800493240356
test_accuracy 0.7952380952380953
test_precision_weighted 0.808770037832538
test_recall_weighted 0.7952380952380953
test_f1_weighted 0.793228612782052
```

طبقه بند SVM:

در این طبقه بند نیز به ازای c های مختلف و همچنین kernel های مختلف تمام مراحل آموزش و اعتبار سنجی را انجام دادیم و بهترین نتیجه مطابق شکل زیر است. البته حالتی که از standard scaler استفاده نشود را نیز بررسی کردیم و متوجه شدیم نتایج در این حالت بهتر است.

```
best accuracy of svm with 8-hold cross validation is:
77.64880952380952 %
hyper parameter: c = 1
hyper parameter: kernel = linear
-----
fit_time 0.8237262070178986
score_time 0.016640156507492065
test_accuracy 0.7764880952380953
test_precision_weighted 0.8177944624819624
test_recall_weighted 0.7764880952380953
test_f1_weighted 0.7682338494838494
```

طبقه بند Decision Tree:

تمام مراحل فوق را این بار به ازای max depth ها مختلف تست کرده و نتایج به شکل زیر است.

```
best accuracy of decision tree with 8-fold cross validation is:
78.86904761904762 %
hyper parameter: max depth = 12
-----
fit_time 0.013138800859451294
score_time 0.013396143913269043
test_accuracy 0.7886904761904762
test_precision_weighted 0.8041215959965959
test_recall_weighted 0.7886904761904762
test_f1_weighted 0.7835983874558423
```

طبقه بند KNN:

تمام مراحل فوق را این بار به ازای n_neighbors های مختلف تست کرده و نتایج به شکل زیر است.

```
best accuracy of KNN with 8-fold cross validation is:
83.18452380952381 %
hyper parameter: n_neighbors = 1
-----
fit_time 0.033054620027542114
score_time 0.03331857919692993
test_accuracy 0.8318452380952381
test_precision_weighted 0.8359424603174603
test_recall_weighted 0.8318452380952381
test_f1_weighted 0.8314840130629604
```

طبقه بند Random forest regression:

تمام مراحل فوق را این بار به ازای max depth ها مختلف تست کرده و نتایج به شکل زیر است.

```
best accuracy of random forest with 8-fold cross validation is:
```

```
77.64880952380952 %
```

```
hyper parameter: max_depth = 3
```

```
-----  
fit_time 0.1870291829109192
```

```
score_time 0.02338549494743347
```

```
test_accuracy 0.7764880952380953
```

```
test_precision_weighted 0.7925550144300145
```

```
test_recall_weighted 0.7764880952380953
```

```
test_f1_weighted 0.7731552346338993
```

طبقه بند perceptron:

```
best accuracy of perceptron with 8-fold cross validation is:
```

```
69.58333333333333 %
```

```
-----  
fit_time 0.013980001211166382
```

```
score_time 0.010817408561706543
```

```
test_accuracy 0.6958333333333333
```

```
test_precision_weighted 0.7355304913340628
```

```
test_recall_weighted 0.6958333333333333
```

```
test_f1_weighted 0.6853373267543243
```

طبقه بند mlpclassifier:

```
best accuracy of MLPClassifier with 8-fold cross validation is:
```

```
76.42857142857142 %
```

```
hyper parameter: hidden_layer_sizes = (50,)
```

```
-----  
fit_time 0.3862118721008301
```

```
score_time 0.01312604546546936
```

```
test_accuracy 0.7642857142857142
```

```
test_precision_weighted 0.779718475968476
```

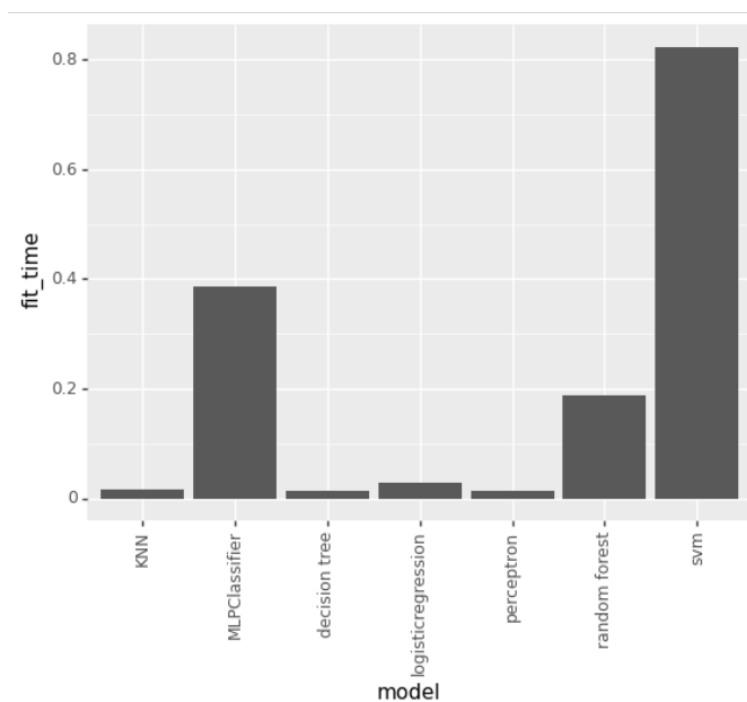
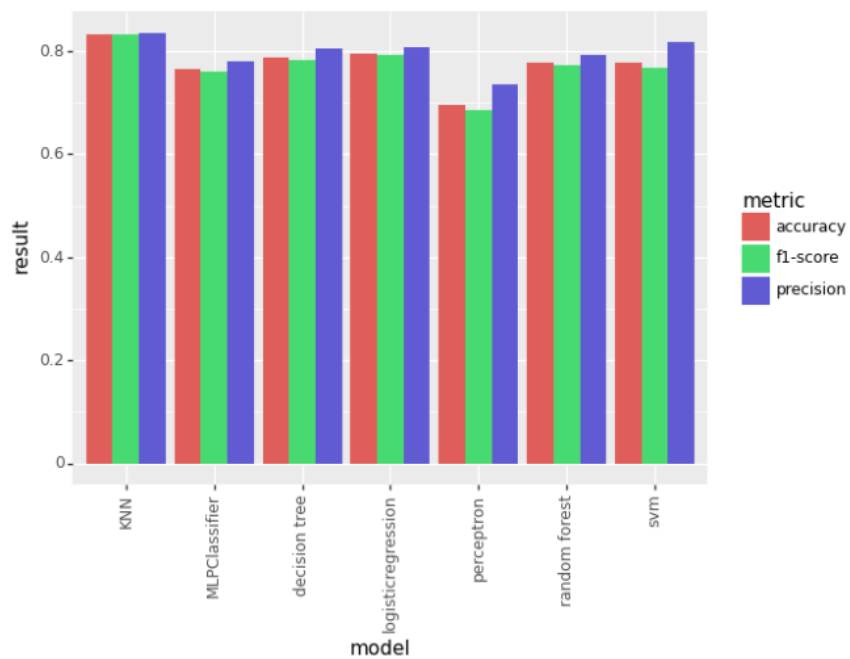
```
test_recall_weighted 0.7642857142857142
```

```
test_f1_weighted 0.7614266408144286
```

به ازای تعداد لایه مخفی ۱، ۲، ۳ و تعداد نود های مختلف آن را بررسی کردم.

جمع بندی:

در نهایت متریک های بدست آمده را در نمودار زیر برای مدل های مختلف ملاحظه می کنید.



نتایج نشان می‌دهد مدل knn بهترین نتیجه را در سه متریک accuracy, f1-score, precision بدست آورده است. همچنین در نمودار آخر ملاحظه می‌کنید زمان آموزش svm از همه بیشتر بوده است.