سوال اول

فایل های این سوال در پوشه ی Source Codes با عنوان P1 قرارداده شده است.

بخش ۱

از کتابخانه های ذیل برای انجام امور استفاده نمودیم. (مراحل این بخش در فایل P1_crawling.ipynb قرار داده شده است.)

```
from bs4 import BeautifulSoup
import requests
import time
```

در ابتدا اطلاعات مربوط به صفحه ی خودروی ۲۰۶ تیپ دو را به صورت دستی از سایت دیوار بیرون کشیدیم، سپس دریافتیم که در هر صفحه صرفا ۲۴ مورد فایل وجود دارد و برای لود شدن باقی موارد نیز به اسکرول است، پس کتابخانه های خزشگر این قابلیت را نداشتند از این رو APIی که صدا زده می شد برای قدم بعدی را بررسی نمودیم و با استفاده از آن در ابتدا یک مجموعه دادگانی تحت عنوان Tokens.txt ساختیم که توکن مربوط به هر محصول در دیوار بود که این توکن در انتهای آدرس هر مورد قرار می گرفت و اجازه دسترسی را برای ما بوجود می آورد.

```
    url = 'https://divar.ir/v/-/{token}'.format(token=token)
```

حال لازم است تا توکن های استخراج شده (۴۰۰۰ مورد توکن) را داخل لینک فوق قرار داده و اطلاعات مورد نیاز هر خودرو را بیرون بکشیم، از ابتدا می دانستیم که اجرای این دستورات و ارسال درخواست های زیاد به سایت دیوار موجب می شود تا بخشی از داده ها را برای ما نفرستد از این رو در برخی مقاطع جهت اعمال فعالیت ها از sleep استفاده نمودیم.

از ۴۰۰۰ توکنی که داشتیم توانستیم در نهایت اطلاعات ۱۰۰۰ خودرو را با توجه به وقت محدود از سایت دریافت کنیم که این تعداد نیز طی چندین مرحله پیاپی اتفاق افتاد و صرفا به این تعداد بسنده کردیم چرا که مشخصا انجام کار به صورت کلی حائز اهمیت است و صرف زمان بیشتر می تواند مشکلات را رفع نماید. (این فرایند در چند مرحله انجام شد که اخرین مرحله ۸۵ دقیقه زمان برد)

تصویری از دیتاست نهایی که ساخته شد در ادامه قرار داده شده است.

	توکن	کارکرد	مدل (سال تولید)	رنگ	نوع سوخت	وضعیت موتور	وضعیت شاسی‌ها	وضعیت بدنه	مهلت بیمهٔ شخص ثالث	گیربکس	قیمت فروش نقدی	نوع آگهی	مایل به معاوضه	شاسی جلو	حداقل مبلغ پیش‌پرداخت	مبلغ هر قسط	تعداد اقساط	شاسی عقب
0	AZdO0Ehe	۲۴۰,۰۰۰	۱۳۸۷	نقره‌ای	بنزینی	سالم	سالم و پلمپ	خط و خش جزیی	ماه ۶	دنده‌ای	۲۴۰,۰۰۰,۰۰۰ تومان	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	AZdKkX0Z	۱۰۵,۰۰۰	۱۳۹۰	سفید	بنزینی	سالم	سالم و پلمپ	رنگ‌شدگی	ماه ۱۰	دنده‌ای	۳۰۳,۰۰۰,۰۰۰ تومان	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	AZdG0oTg	۶,۸۰۰	۱۴۰۰	نوک‌مدادی	بنزینی	سالم	NaN	سالم و بی‌خط و خش	ماه ۶	دنده‌ای	۴۸۰,۰۰۰,۰۰۰ تومان	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	AZ0IEKbL	۳۸,۸۰۰	۱۳۹۹	سفید	بنزینی	سالم	سالم و پلمپ	سالم و بی‌خط و خش	ماه ۳	دنده‌ای	۴۳۲,۰۰۰,۰۰۰ تومان	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	AZROGMH3	۱۲۴,۰۰۰	۱۴۰۱	سفید	بنزینی	سالم	سالم و پلمپ	سالم و بی‌خط و خش	ماه ۵	دنده‌ای	۴۶۹,۰۰۰,۰۰۰ تومان	فروشی	NaN	NaN	NaN	NaN	NaN	NaN

بخش ۲

ما در بخش قبلی ۱۰۰۰ توکن از ۴۰۰۰ توکنی که داشتیم را اطلاعات اساسی اش را بررسی کردیم و در یک فایل با عنوان output.csv ذخیره نمودیم، برای اضافه کردن اطلاعات هر توکن مانند تصویر و توضیحات نیز قصد داریم تا بخشی از داده ها را استفاده کنیم چرا که کار زمان بر و با حجم بالایی می شود و از طرفی از سمت سایت دیوار محدودیت بیشتری قطعا اعمال می گردد.

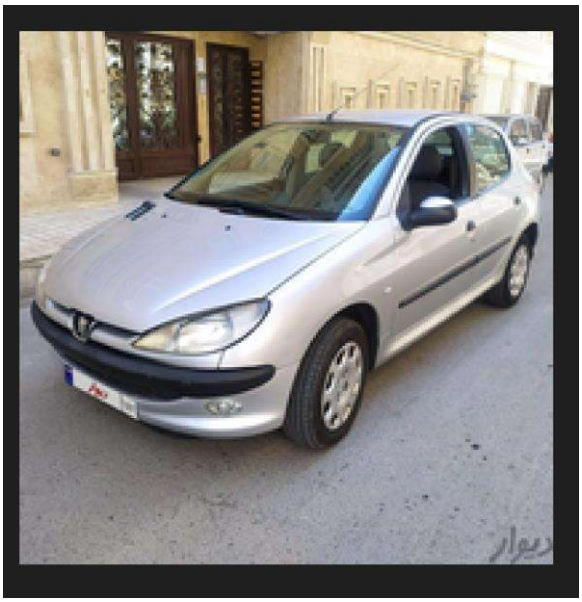
(مراحل این بخش در فایل P1_2.ipynb قرار داده شده است.)

در قدم اول اقدام به تمیز کردن فایل دیتاست نمودیم سپس دو قطعه کد توسعه داده شد تا توضیحات فارسی مربوط به اطلاعات هرماشین و همچنین تصویر اصلی هر مورد به دیتاست افزوده شود.

دیتاست نهایی :

	تصویر	توضیحات	قیمت فروش نقدی	گیربکس	مهلت بیمهٔ شخص ثالث	وضعیت بدنه	وضعیت شاسی‌ها	وضعیت موتور	نوع سوخت	رنگ	مدل (سال تولید)	کارکرد	توکن
575	b''	بسیار سالم کم کارکرد به شرط داخل کاب\n\nکارشناسی...	۲۴۰,۰۰۰,۰۰۰ تومان	دنده‌ای	ماه ۱۲	رنگ‌شدگی	سالم و پلمپ	سالم	بنزینی	نقره‌ای	۱۳۸۷	۱۴۷,۰۰۰	AZa66TzW
410	b''	بدنه بدون رنگ هست و مقداری خط و خش جزیی دارد ...	۲۶۰,۰۰۰,۰۰۰ تومان	دنده‌ای	ماه ۲	خط و خش جزیی	سالم و پلمپ	سالم	بنزینی	سفید صدفی	۱۳۸۷	۲۰۰,۰۰۰	AZbODON2
330	b''	سلام. ماشین چند ماه هست که داخل پارکینگ مستقفه...	تومان ۱۱,۲۱۱,۱۱۱	دنده‌ای	ماه ۸	سالم و بی‌خط و خش	سالم و پلمپ	سالم	بنزینی	سفید	۱۴۰۱	۱,۶۰۰	AZbSO_0P
593	b''	دو لک رنگ دارد فنی سالم سند تک برگ	۳۷۰,۰۰۰,۰۰۰ تومان	دنده‌ای	ماه ۸	رنگ‌شدگی	سالم و پلمپ	سالم	بنزینی	سفید صدفی	۱۳۹۸	۶۰,۰۰۰	AZayprhc
795	b''	تیپ دو کاملا سالم بیرنگ بدون۲۰۶ ... تصادف کیلومتر	۴۴۰,۰۰۰,۰۰۰ تومان	دنده‌ای	ماه ۶	سالم و بی‌خط و خش	سالم و پلمپ	سالم	بنزینی	خاکستری	۱۳۹۸	۵۰,۰۰۰	AZZGpbta

یک نمونه تصویر خروجی از هر مورد :



تصویر ارائه شده مربوط به اولین ردیف از دیتاست نهایی می باشد که در بخش فوقانی قرار داده شده است.

بخش ۳

اقدامات زیادی روی داده ها در این مرحله انجام شد، به طوریکه تمامی مراحل پیش پردازش پیاده سازی شد.

- داده هایی که صفر بودند پر شدند.
- اعداد فارسی به اعداد انگلیسی تغییر کرد.
- پسوند های اضافه مانند تومان و ماه حذف شد
- داده های عددی به عدد تبدیل شدند.
- سر ستون ها تغییر نام داده شدند.
- ستون های بدون استفاده مانند توکن و نوع سوخت حذف شدند.
- عملیات لیبل زدن داده های غیر عددی انجام شد و لیبل ها نیز درفایل ذخیره شد.
- استاندارد سازی انجام شد و میانگین ها و انحراف از معیار ها در فایل ذخیره شد.
- مدل رگرسیون خطی ساخته شد و مدل نیز ذخیره شد.

	Usage	Year	Color	Engine	Chassis	Body	Insurance	Gearbox	Price
0	240000.0	1387	17	1	1	2	6	1	240.0
1	105000.0	1390	10	1	1	4	10	1	303.0
2	6800.0	1400	18	1	1	7	6	1	480.0
3	38800.0	1399	10	1	1	7	3	1	432.0
4	24000.0	1401	10	1	1	7	5	1	469.0
5	16000.0	1400	5	1	1	7	10	1	455.0
6	228000.0	1387	10	1	1	7	8	1	260.0
7	117000.0	1396	10	1	1	2	3	1	365.0
8	252000.0	1389	17	1	1	4	6	1	265.0
9	107000.0	1385	3	1	1	7	9	1	285.0

(مراحل این بخش و بخش ۴ در فایل P1_3.ipynb قرار داده شده است.)

بخش ۴

خروجی نهایی معیار های مدل:

```
R-squared: 0.6205064152622783
Mean Squared Error: 4759.280698979096
Mean Absolute Error: 38.36179018644281
```

بخش ۵

با توجه به اینکه نیاز بود یک اسکریپت نوشته شود و خارج از فایل اصلی بودیم نیاز داشتیم تا داده های لیبل زنی و معیارهای میانگین و انحراف از معیار را ابتدا از فایل بخوانیم، سپس مدل را از فایل بخوانیم.

برای انجام عملیات لازم است تا ورودی را از کاربر بگیریم و در نهایت خروجی را محاسبه کنیم.

(مراحل این بخش در فایل P1_5.ipynb قرار داده شده است.)

گرفتن ورودی از کاربر:

```
0 : * آبی *      * آلبالویی * 1 :      * بادمجانی * 2 :      * بژ * 3 :      * نیتانیوم * 4 :      * خاکستری * 5 :      * دلفینی * 6 :      * ذغالی * 7 :
--------------------
0 : * تعویض شده *      * سالم * 1 :    * نیاز به تعمیر * 2 :  *  *
--------------------
0 : * رنگشده *    * سالم و پلمپ * 1 :    * ضربه‌خورده * 2 :  *  *
--------------------
0 : * تصادفی *    * تمام‌رنگ * 1 :      * خط و خش جزیی * 2 :    * دوررنگ * 3 :    * رنگشدگی * 4 :      * رنگشدگی، در 1 ناحیه * 5 :    * رنگشدگی، در 2 ناحیه *
--------------------
0 : * دنده‌ای اتوماتیک * 1 :  *  *  *
--------------------
[240000, 1387, 17, 1, 1, 2, 6, 1]
[ 2.64570824 -0.36301818 -0.38029283 -0.38049458 -0.38049458 -0.38048197
 -0.38043153 -0.38049458]
```

خروجی:

```
array([226.63665453])
```

عددی که باید پیش بینی میشد :

Usage	Year	Color	Engine	Chassis	Body	Insurance	Gearbox	Price
240000.0	1387	17	1	1	2	6	1	240.0

نتیجه نهایی خروجی قابل قبولی دارد.

تست نهایی بر اساس داده های آموزش ندیده سایت دیوار:

خروجی برنامه :

```
price prediction :  [226.80667661]
```

قیمت و نظر داخل دیوار :

۲۷۰ میلیون تومان