به نام خدا

پروژه نهایی

آمار و احتمالات مهندسی

استاد: دکتر ابوالقاسمی

عرفان پناهی

۸۱۰۱۹۸۳۶۹

نیمسال اول ۰۰-۹۹

فهرست مطالب:

توضیحات مربوط به فایل کد های پیوست شده:

در قسمت های مورد نیاز در گزارش پروژه از اسکریپتا پایتون استفاده شده که در همان قسمت نیز بخش مورد نظر در فایل CA_code_810198369 مشخص شده است.

۱ ضریب اویلر - ماسکرونی

قسمت آ) با توجه به اینکه n عددی است که آنرا در ابتدا اختیار میکنیم و سپس آنرا بع اعداد صحیح کوچکتر و مساوی خودش تقسیم میکنیم، پس مقدار ثابتی خواهد داشت و نمیتواند متغیر تصادفی باشد. از سوی دیگر r همواره در حال تغییر است و مقادیر ۱ تا n را اختیار میکند و بنابراین متغیرتصادفی مسئله r است.

قسمت ب) مقدار ε_r (خطا) به ازای هر r مشخص بصورت «اختلاف سقف خاج قسمت تقسیم و مقدار دقیق آن» تعریف میشود. روابط ریاضی آن بصورت زیر خواهد بود:

$$\varepsilon_r = \left|\left\lceil\frac{n}{r}\right\rceil - \frac{n}{r}\right| \rightarrow \left\lceil\frac{n}{r}\right\rceil \geq \frac{n}{r} \rightarrow \varepsilon_r = \left\lceil\frac{n}{r}\right\rceil - \frac{n}{r}$$

از طرفی r خود توزیعی یکنواخت گسسته از ۱ تا n است. پس تابع احتمال این توزیع بصورت زیر خواهد بود:

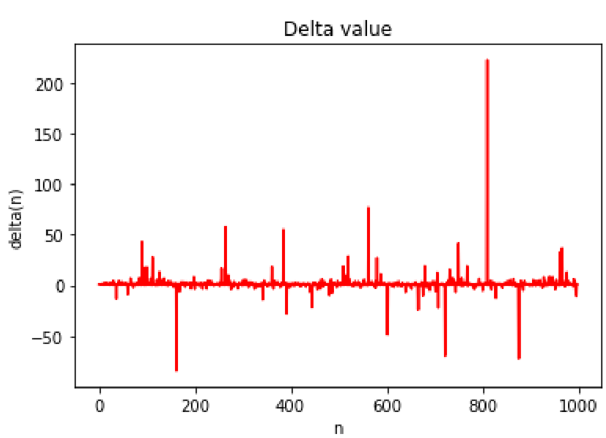
$$P_R(r) = \frac{1}{n}$$

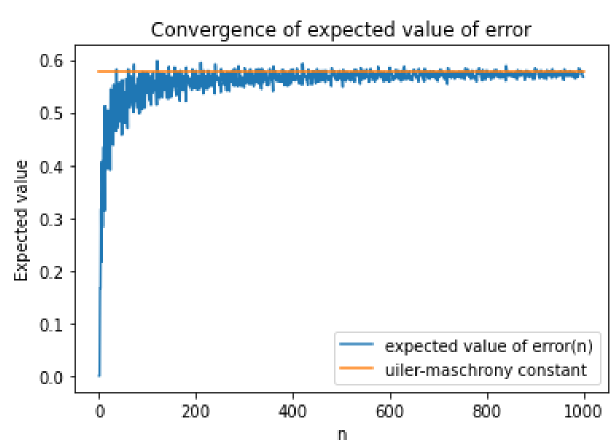
حال میخواهیم امید ریاضی خطا را نیز محاسبه کنیم که بکمک رابطه زیر میتوان آنرا محاسبه نمود:

$$E(\varepsilon_r) = \sum_{r=1}^{n} \varepsilon_r P_R(r) = \sum_{r=1}^{n} \left(\left\lceil\frac{n}{r}\right\rceil - \frac{n}{r}\right)\frac{1}{n}$$

حال این امید را به کمک اسکریپت پایتون Part a قسمت ۱ محاسبه میکنیم.

با توجه به اینکه مقدار n باید یک عدد بزرگ باشد به ازای $n = 1000$ مقدار امید ریاضی خطا را محاسبه میکنیم. گزارش اثبات همگرایی امید و همچنین نمودار δ_n نیز بصورت زیر خواهد بود:(Part b قسمت ۱)





۲ مسألهی نیوتن-پیپس

قسمت آ) را با A نشان ابتدا برای بررسی منظمانه نام گذاری هایی برای پیشامدها انجام میدهیم:

پیشامد A: در پرتاب شش تاس، حداقل یکی از تاس ها ۶ باشد.

پیشامد B: حداقل دو تاس از ۱۲ تاس پرتاب شده ۶ باشد.

پیشامد C: حداقل سه تاس از ۱۸ تاس پرتاب شده ۶ باشد.

پاسخ نیوتن به پیپس بصورت زیر بود:

« از طریق اصل متمم باید احتمال A حساب شود. برای اینکار تعداد حالاتی که هرگز تاس ۶ نیاید را از کل حالات کم میکنیم. حال برای پیشامد های B و C به ترتیب تعداد پرتاب ها را به ۲ و ۳ گروه شش تایی تقسیم میکنیم و به همین ترتیب احتمال های مربوط به پیشامد های B و C محاسبه میشود. »

مشکل این پاسخ اینجاست که حالاتی که در بین هردسته بندی دوبار (یا سه بار) تاس ۶ بیاید درنظر گرفته نشده است. به همین خاطر سعی میکنیم راه حلی جدید برای این مسئله معرفی کنیم که تمامی حالات در نظر گرفته شود:

راه حل: برای حل این مسئله از اصل متمم استفاده میکنیم. حالات نامطلوب حالاتی است که تعداد دفعاتی که تاس شش میآید از حداقل موردنظر کمتر باشد. برای پیشامد A حالات نامطلوب زمانی است که ۰ تاس از ۶ بار پرتاب شش بیاید. برای پیشامد B حالات نامطلوب زمانهایی است که ۰ یا ۱ تاس از ۱۲ تا پرتاب شش بیاید. حالات نامطلوب پیشامد C زمانی است که ۰ ، ۱ یا ۲ تاس از بین ۱۸ تاس شش بیاید.

از سوی دیگر حالات نامطلوب به نوعی توزیع دوجمله ای است.

توزیع دوجمله ای: متغیرتصادفی X بیانگر تعداد موفقیت در n بار پرتاب است.

$$X \sim Bin(n,p) \qquad P_X\{X=k\} = \binom{n}{k} p^k (1-p)^{n-k}$$

در مسئله ما p (احتمال موفقیت) برابر احتمال ۶ آمدن یک تاس سالم یعنی $\frac{1}{6}$ است. پس برای هر پیشامد احتمال مربوطه را محاسبه مینماییم:

$$P(A) = 1 - P_X\{X=0\} = 1 - \left(\frac{5}{6}\right)^6 = 0.6551$$

$$P(B) = 1 - \sum_{i=0}^{1} P_X\{X=i\} = 1 - \sum_{i=0}^{1} \binom{12}{i}\left(\frac{1}{6}\right)^i\left(\frac{5}{6}\right)^{12-i} = 0.6187$$

$$P(B) = 1 - \sum_{i=0}^{2} P_X\{X=i\} = 1 - \sum_{i=0}^{2} \binom{18}{i}\left(\frac{1}{6}\right)^i\left(\frac{5}{6}\right)^{18-i} = 0.5973$$

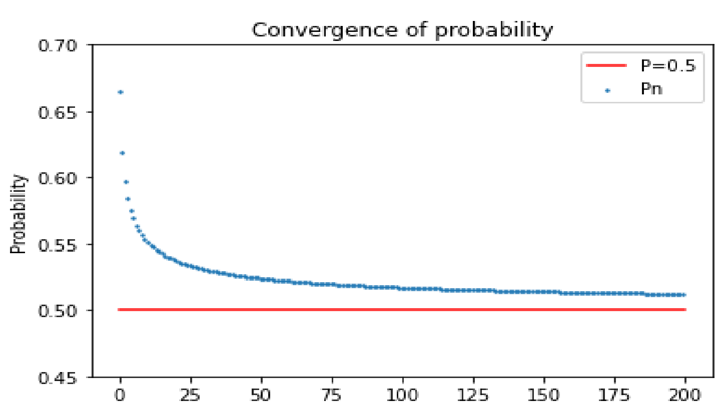
قسمت ب) حال به سادگی میتوانیم تعمیم را برای مشاهده حداقل n تا شش در میان $6n$ تا پرتاب محاسبه کنیم.

$$P_n = 1 - \sum_{i=0}^{n-1} P_X\{X=i\} = 1 - \sum_{i=0}^{n-1} \binom{6n}{i}\left(\frac{1}{6}\right)^i\left(\frac{5}{6}\right)^{6n-i}$$

$$\rightarrow P_n = 1 - \sum_{i=0}^{n-1} \frac{(6n)!}{(6n-i)!\,i!}\left(\frac{1}{6}\right)^i\left(\frac{5}{6}\right)^{6n-i}$$

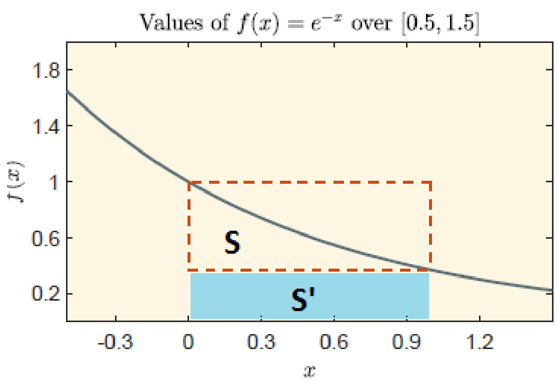
اثبات همگرایی: به کمک اسکریپت پایتون Part b قسمت ۲ همگرایی را برای n های مختلف تا $n = 1000$ محاسبه می‌کنیم. همچنین مقدار P_n را نیز به ازای $n = 1000$ محاسبه می‌کنیم.

نمودار همگرایی P_n بصورت زیر خواهد بود:



۳ تخمین عدد نپرین با روش مونته کارلو

قسمت آ) طبق نمودار داده شده با توجه به داده های زیر مساحت مستطیل جدا شده را محاسبه میکنیم:



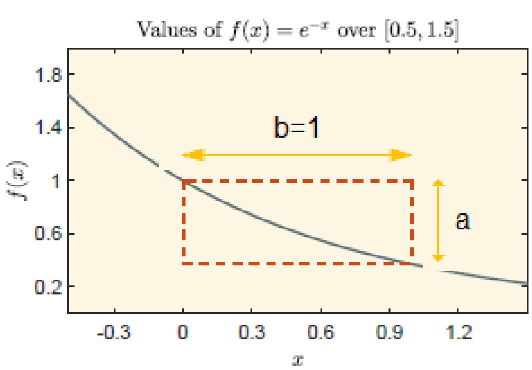
باتوجه به اینکه $I = S + S'$ در شکل روبه رو بیانگر انتگرال (مساحت زیر منحنی) e^{-x} است. پس با محاسبه I و رابطه $S = I - S'$ میتوان به خواسته مسئله رسید.

$$S = I - S' = \int_0^1 e^{-x}\,dx - e^{-1} = 1 - 2e^{-1}$$

قسمت ب) طبق رابطه محاسبه شده برای مساحت حال e را بر حسب S مینویسیم:

$$S = 1 - 2e^{-1} \rightarrow 2e^{-1} = 1 - S \rightarrow e^{-1} = \frac{1 - S}{2} \rightarrow e = \frac{2}{1 - S}$$

قسمت ج) ابتدا باید S_{rect} را محاسبه نماییم:



طول مستطیل : $b = 1$

عرض مستطیل : $a = e^0 - e^{-1} = 1 - e^{-1}$

پس مساحت متطیل داده شده بصورت زیر است.

$$S_{rect} = ab = 1 - e^{-1} \rightarrow S_{rect} = \frac{S + 1}{2}$$

در نتیجه سمت چپ رابطه گفته شده بصورت زیر ساده تر میشود:

$$\frac{S}{S_{rect}} = \frac{2S}{S + 1} = \lim_{n \to \infty} \frac{n_{red}}{n} = P \rightarrow S = \frac{P}{2 - P}$$

حال ثابت اویلر را به کمک اسکریپت پایتون Part c قسمت ۳ تخمین میزنیم.

خروجی کد پس از ران کردن بصورت زیر شده است:

```
The estimation of the Euler constant is : 2.7241379310344827
```

قسمت د) برای یک توزیع نرمال تخمین بیشینه درست نمایی میانگین و واریانس بصورت زیر است:

$$f_{\mathbb{X}}(\mathbb{X};\mu,\sigma) = \prod_{i=1}^{n} f_X(X_i;\mu,\sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\Pi}} e^{-\frac{(x_i^2-\mu)}{2\sigma^2}}$$

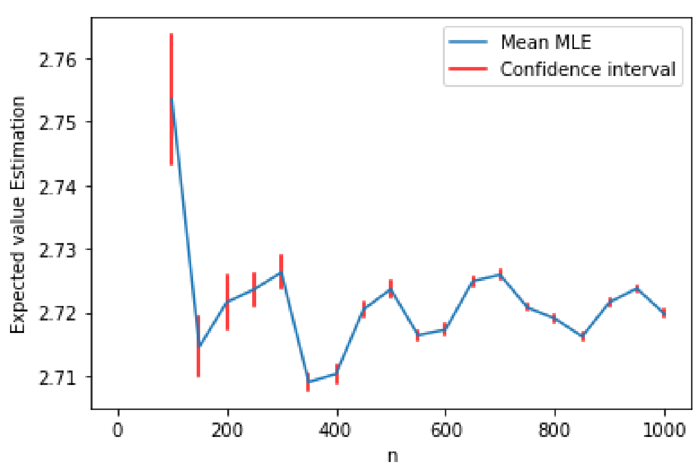
$$Ln\big(f_{\mathbb{X}}(\mathbb{X};\mu,\sigma)\big) = \sum_{i=1}^{n} Ln\big(f_X(X_i;\mu,\sigma)\big) = \sum_{i=1}^{n}\left(-Ln\big(\sigma\sqrt{2\Pi}\big) - \frac{(x_i^2-\mu)}{2\sigma^2}\right)$$

$$MLE\ for\ \mu \rightarrow \frac{\partial}{\partial\mu}\Big[Ln\big(f_{\mathbb{X}}(\mathbb{X};\mu,\sigma)\big)\Big] = 0 \rightarrow \hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$MLE\ for\ \sigma \rightarrow \frac{\partial}{\partial\sigma}\Big[Ln\big(f_{\mathbb{X}}(\mathbb{X};\mu,\sigma)\big)\Big] = 0 \rightarrow \hat{\sigma}_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu}_{ML})^2$$

حال به کمک این روابط میانگین و واریانس به ازای $n = 100$ ، را به کمک اسکریپت پایتون Part d قسمت ۳ تخمین میزنیم.

قسمت ه) نمودار امید برحسب n که کد آن به کمک اسکریپت پایتون Part e قسمت ۳ رسم شده است بصورت زیر خواهد بود.



۴ سری فیبوناچی تصادفی

قسمت آ) بازه اطمینان ۹۵٪ به کمک روابط زیر بدست خواهد آمد:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - X_n)^2$$

$$Z_{1-\frac{\alpha}{2}} = Z_{1-\frac{5\%}{2}} = Z_{97.5\%} = 1.96$$

$$Confidence\ Interval: \bar{X}_n \pm \frac{S_n . Z_{1-\frac{\alpha}{2}}}{\sqrt{n}}$$

حال با استفاده از اسکریپت پایتون Part a قسمت ۴ سعی میکنیم بازه اطمینان را برای سه تعداد نمونه ۱۰۰، ۲۰۰ و ۳۰۰ تخمین بزنیم. خروجی این اسکریپت به صورت زیر خواهد بود:

```
Confidence interval for 100 samples: (1.0539953376005309,1.1342811795003558)
Confidence interval for 200 samples: (1.1122695860505571,1.1395618784678778)
Confidence interval for 300 samples: (1.1156848496805154,1.1348770232804968)
```

قسمت ب) رابطه تخمین بیشینه بزرگنمایی پارامتر توزیع نمایی (λ) بصورت زیر خواهد بود:

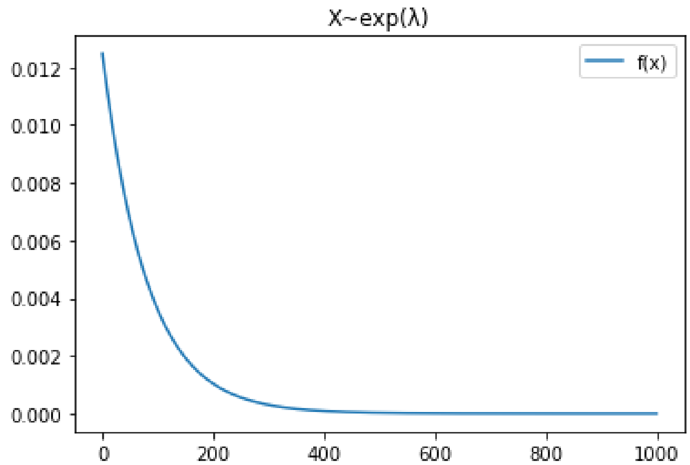
$$X \sim \exp(\lambda) \to f_X(x) = \lambda e^{-\lambda x} \to f_{\mathbb{X}}(\mathbb{X};\lambda) = \prod_{i=1}^{n} f_X(X_i;\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

$$Ln(f_{\mathbb{X}}(\mathbb{X};\lambda) = \prod_{i=1}^{n} Ln\left(\lambda e^{-\lambda x_i}\right) = n(\ln(\lambda)) - \lambda \sum_{i=1}^{n} x_i = LL(\lambda)$$

$$\frac{\partial LL(\lambda)}{\partial \lambda} = 0 \to \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \to \hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$$

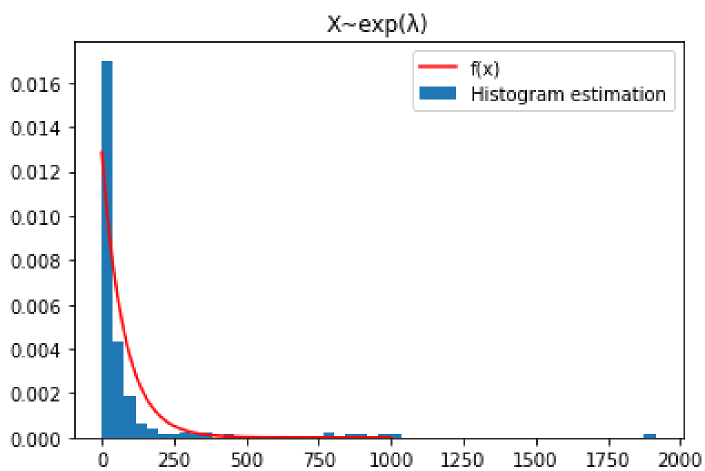
حال به کمک اسکریپت پایتون Part b قسمت ۴ با ۱۰۰۰ نمونه (ماکزیمم جمله در ۲۵ جمله اول دنباله) مقدار پارامتر λ را تخمین بزنیم و تابع چگالی را برازش می‌کنیم. خروجی این کد بصورت زیر خواهد بود:

```
λ = 0.012149339683388207
```



قسمت ج) به کمک اسکریپت پایتون Part c قسمت ۴ با ۲۰۰ نمونه (برای وضوح) خواهیم داشت:

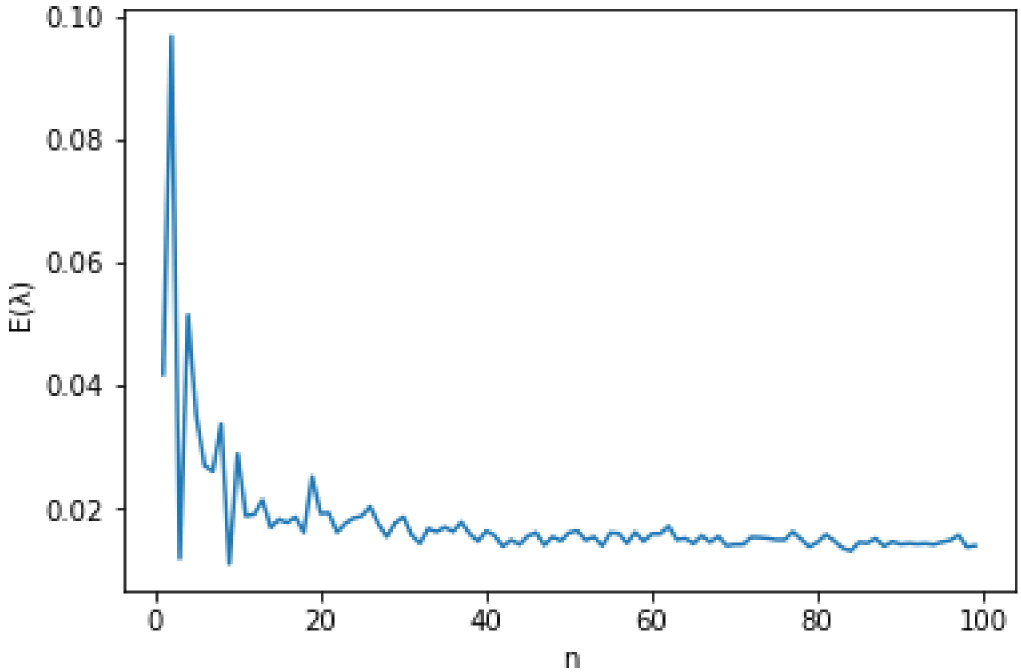
```
λ = 0.012850992739189103
```



قسمت د) برای محاسبه امید سعی میکنیم از ۱۰۰ نمونه لاندا (که با num نشان میدهیم) استفاده می‌کنیم.

$$E(\lambda) = \bar{\lambda}_{num} = \frac{1}{num}\sum_{i=1}^{num} \lambda_i$$

به کمک اسکریپت پایتون Part d قسمت ۴ برای $num = 100$ خواهیم داشت:



۵ قاعدهٔ دایره ای برای مقادیر ویژه

قسمت آ) با توجه به میانگیم و واریانس داده شده سعی میکنیم پارامتر های هر کدام از توزیع های گوسی و یکنواخت را بررسی و برحسب n تعیین کنیم:

توزیع گوسی:

$$\mu = E(X) = 0 \to \mu = 0$$

$$\sigma^2 = E(X^2) - E(X) = \frac{1}{n} - 0 = \frac{1}{n} \to \sigma = \frac{1}{\sqrt{n}}$$

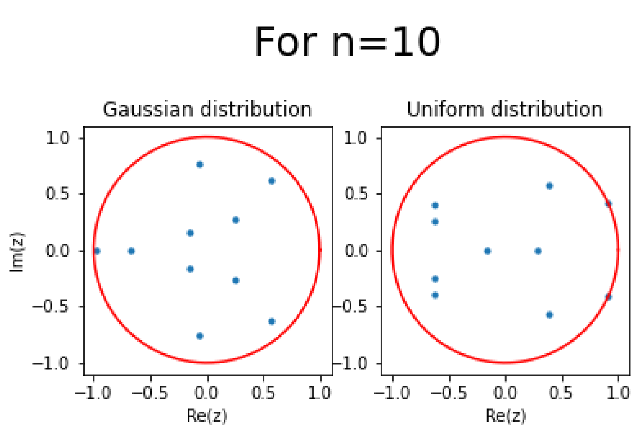
توزیع یکنواخت:

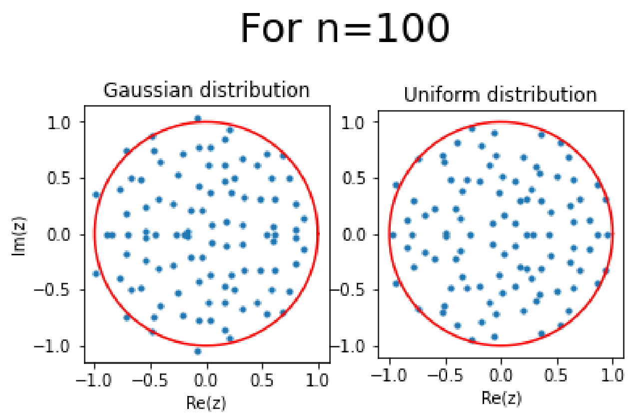
$$\frac{a+b}{2} = E(X) = 0 \to a = -b$$

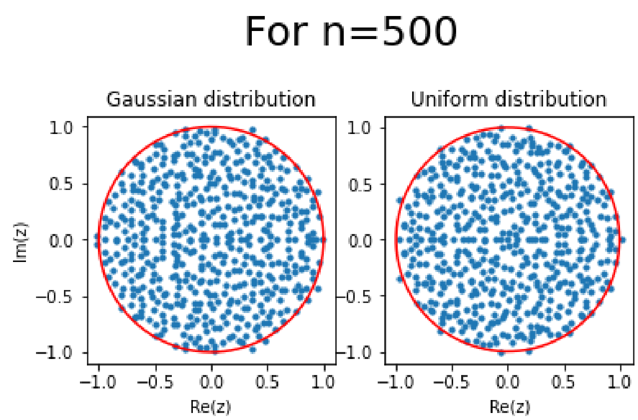
$$\sigma^2 = \frac{(a-b)^2}{12} = \frac{a^2}{3} = \frac{1}{n} \to a = -\sqrt{\frac{3}{n}}\ , b = +\sqrt{\frac{3}{n}}$$

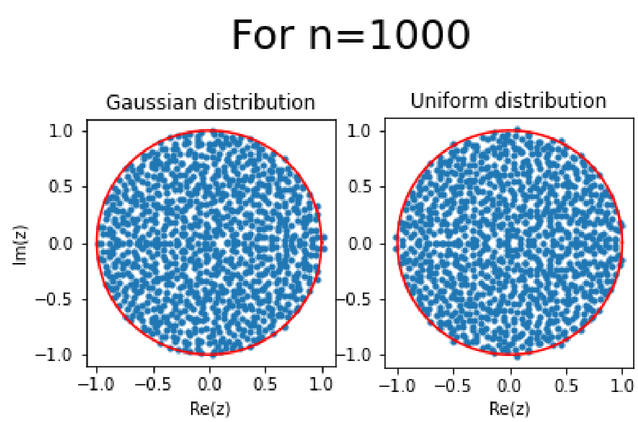
حال با استفاده از این پارامتر ها و اسکریپت پایتون Part a قسمت ۵ با افزایش مقدار n و رسم نمودار ها سعی میکنیم همگرایی مقادیر ویژه در دایره واحد را توجیه کنیم.

نمودار های خروجی برای n=10,100,500,1000 بصورت زیر درآمده است و همانطور که مشاهده میشود با افزایش n تمامی نقاط(مقادیر ویژه) به داخل دایره همگرا میشوند.

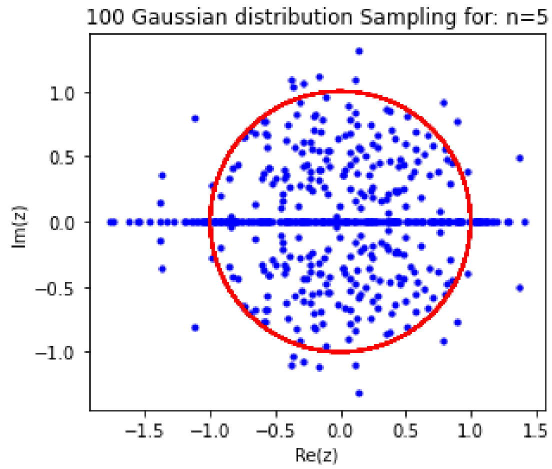


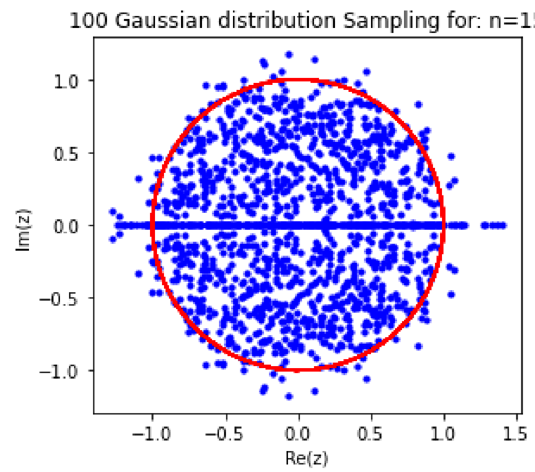


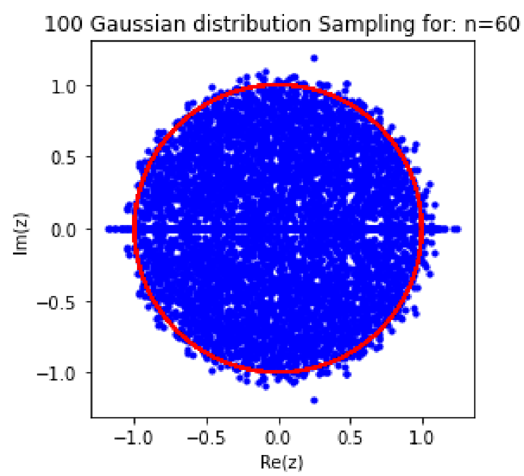




قسمت ب) با استفاده از اسکریپت پایتون Part b قسمت ۵ و با مقادیر n داده شده و رسم نمودار ها سعی میکنیم خواسته مسئله را بررسی کنیم:







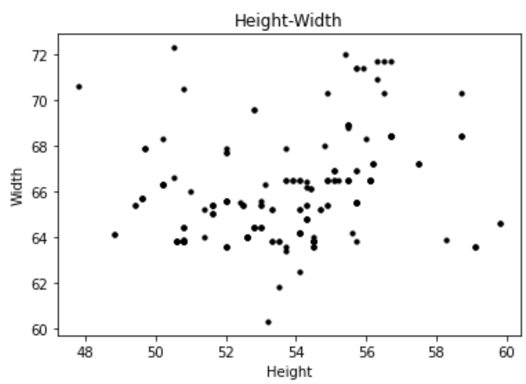
همانطور که مشاهده میشود با افزایش مقدار n تمامی مقادیر ویژه به داخل دایره واحد همگرا میشوند و تقریباً توزیعی مشابه دارند.

۶ دادهٔ بازی

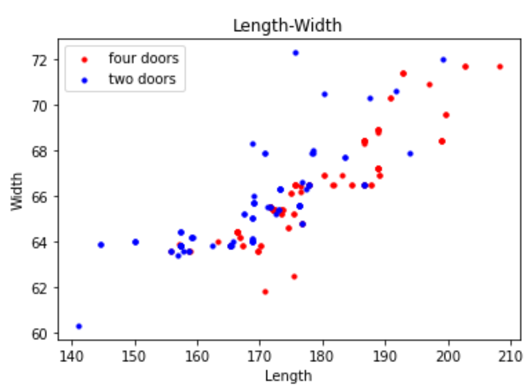
قسمت آ)

آ-آ) برای بررسی این قسمت از اسکریپت پایتون بخش Part a-a قسمت ۶ استفاده میکنیم و با توجه به نمودار های خروجی این اسکریپت فرضیه ها را موشکافی می‌کنیم:

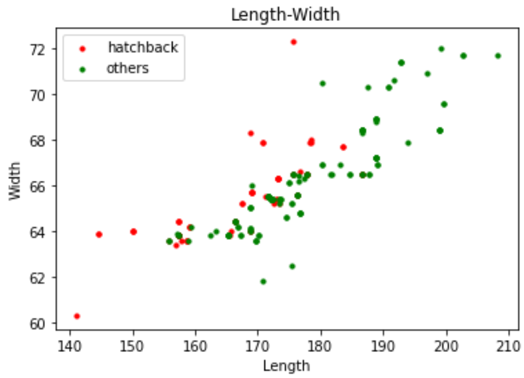
آ-آ-i) همانگونه که در نمودار نیز قابل مشاهده است پراکندگی نقاط تقریباً بصورتی است که ضریب همبستگی زاویه ۴۵ درجه دارد. یعنی به طور معمول و تقریباً با افزایش ارتفاع ماشین، عرض آن نیز افزایش پیدا میکند اما موارد استثنائی نیز وجود دارد. (البته این فرضیه ضعیف است)



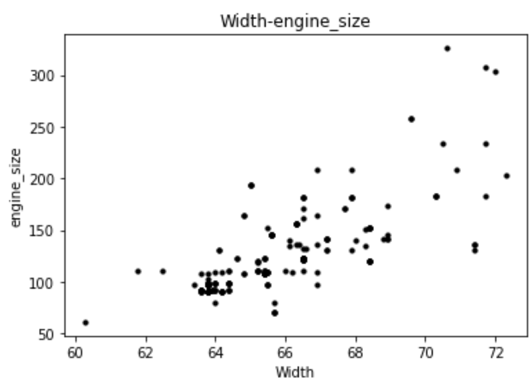
آ-آ-ii) همانطور که در نمودار هم مشاهده میشود نقاط قرمز که مربوط به ماشین های ۴ در هستند در گوشه سمت راست و پایین نمودار قرار میگیرند. یعنی ماشین های ۴ در طول بیشتری دارند.



آ-آ-iii) همانطور که طبق فرض انتظار میرفت ماشین های هاچبک بیشتر خودرو های با کمترین طول و عرض هستند.



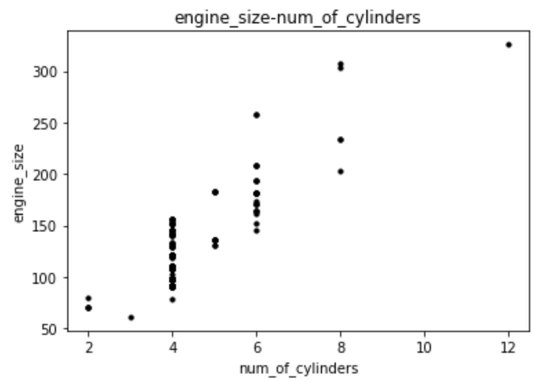
آ-آ-iv) همانطور که انتظار میرفت ظرفیت موتور با عرض ماشین نسبت مستقیم دارد و هرچه خودرو عریض تر باشد حجم موتور بیشتری دارد.



آ-ب) برای بررسی این قسمت از اسکریپت پایتون بخش Part a-b قسمت ۶ استفاده میکنیم و با توجه به نمودار های خروجی این اسکریپت فرضیه ها را موشکافی می‌کنیم:

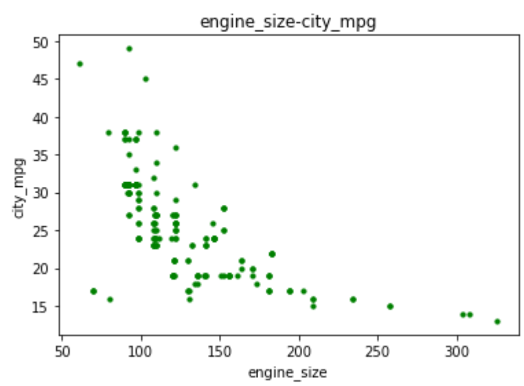
فرض اول: با افزایش تعداد سیلندر در ماشین ها حجم موتور افزایش می‌یابد.

همانطور که در نمودار مشاهده میشود ماشین هایی که تعداد سیلندر بیشتری دارند دارای حجم موتور بیشتری نیز هستند.



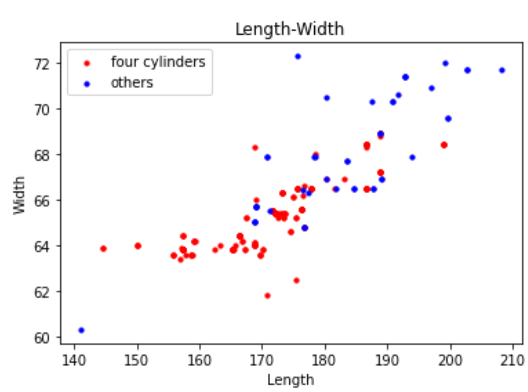
فرض دوم: هرچقدر حجم موتور بیشتر باشد، آلایندگی شهری آن ماشین بیشتر است.

بر خلاف تصور طبق نمودار رسم شده ماشین های با حجم موتور بیشتر میزان آلایندگی کمتری هم دارند.



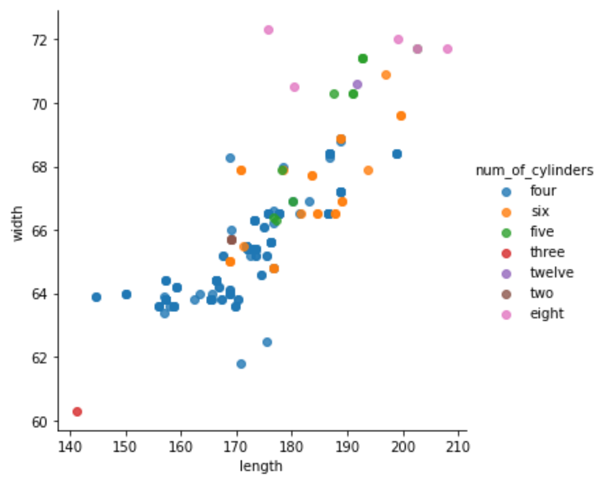
فرض سوم: ماشین های با طول و عرض کمتر تعداد سیلندر های کمتری دارند.

برای بررسی این موضوع ماشین های ۴ سیلندر را انتخاب میکنیم و رسم میکنیم. همانطور که مشاهده میشود این ماشین ها طول و عرض کمتری نسبت به بقیه دارند.



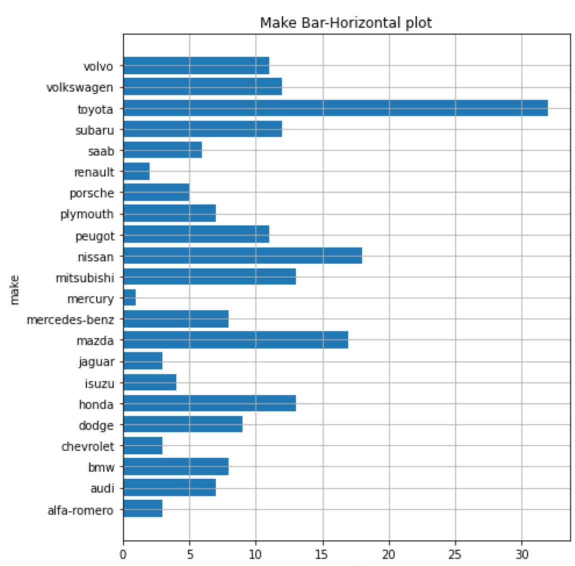
آ-ج) **لغزش** زمانی اتفاق می‌افتد که چندین داده بعد از رسم روی همدیگر قرار گیرند و به همین خاطر نمیتوانیم پراکندگی را در آن نقطه به خوبی مشاهده کنیم. بعبارت دیگر اگر تصویر نمودار نتواند خود واقعی نمودار و پراکندگی در یک نقطه را به خوبی نشان دهد، می‌گوییم **لغزش** اتفاق افتاده است.

برای مثال در رسم نمودار عرض ماشین بر حسب طول، برای تشخیص پراکندگی ماشین های با ۴ سیلندر دچار مشکل میشویم. زیرا این ماشین ها دارای طول و عرض مشابعی هستند برای همین روی همدیگر سوار میشوند. برای رفع این مشکل از اسکریپت پایتون Part a-c قسمت ۶ استفاده میکنیم و نمودار بهبود یافته را رسم میکنیم.



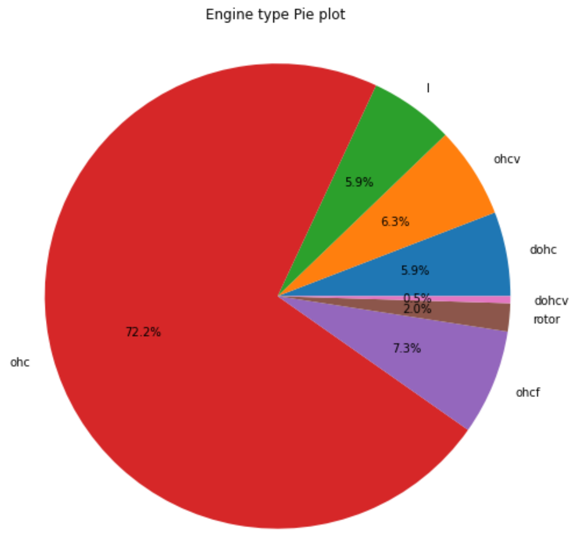
آ-د) ۱. نمودار میله ای افقی سازنده ماشین ها:

همانطور که مشاهده میشود بیشتر ماشین ها مربوط به شرکت تویوتا هستند.



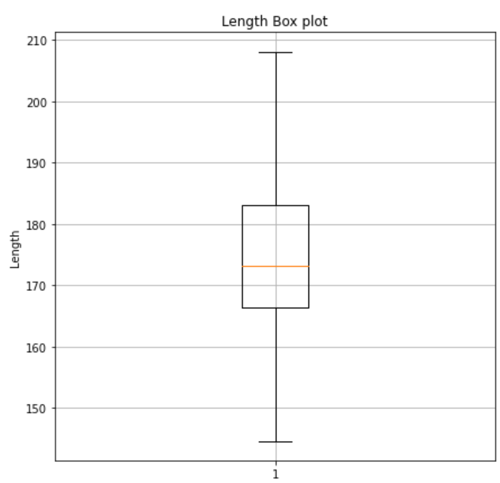
۲. نمودار دایره ای نوع موتور ماشین ها:

همانطور که مشاهده میشود اغلب ماشین ها دارای موتور ohc هستند.



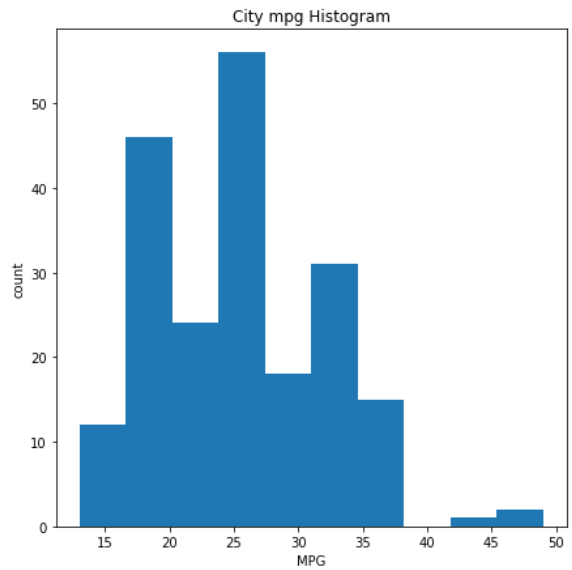
۳. نمودار جعبه ای طول ماشین ها:

همانطور که مشاهده میشود بیشتر ماشین ها دارای طولی بین ۱۶۵ تا ۱۸۵ هستند و میانگین طول در حدود ۱۷۳ است.



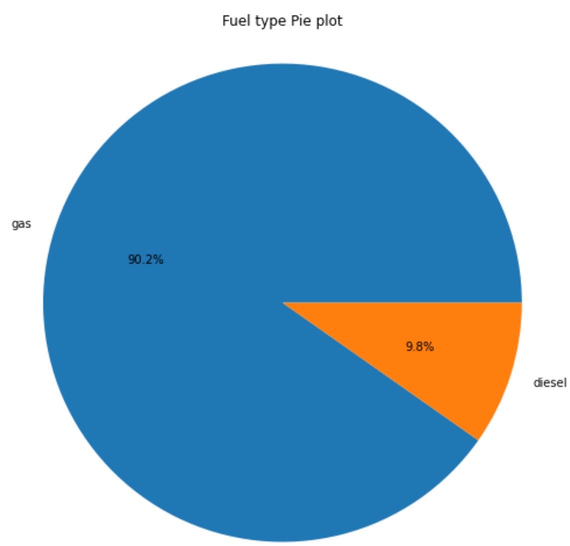
۴. نمودار هیستوگرام میزان آلایندگی شهری:

همانطور که مشاهده میشود بیشتر آلایندگی در حدود ۲۵ است و این میزان توزیعی شبه گوسی دارد.



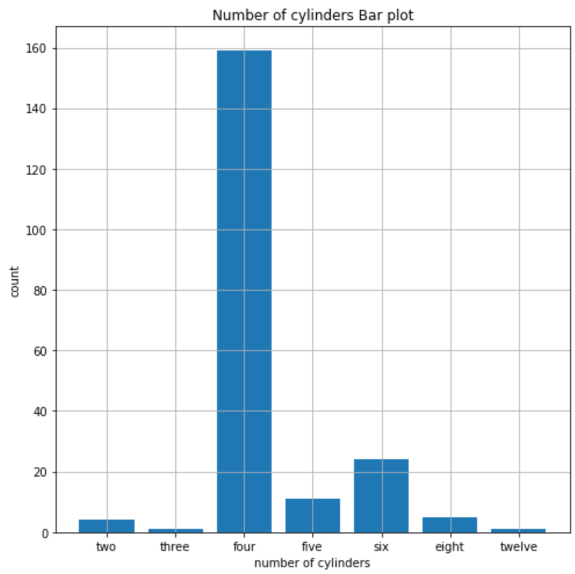
۵. نمودار دایره ای نوع سوخت مصرفی ماشین ها:

همانطور که مشاهده میشود بیشتر ماشین ها سوخت گازی مصرف می‌کنند.



۶. نمودار میله ای تعداد سیلندر ماشین ها:

همانطور که مشاهده میشود بیشتر ماشین ها ۴ سیلندر هستند.



قسمت ب) دو آزمون فرض را بصورت های زیر طراحی می‌کنیم و به کمک اسکریپت پایتون بخش Part b قسمت ۶ نتایج را بررسی می‌کنیم:

آزمون فرض اول: میانگین مصرف سوخت شهری ماشین ها ۲۵ است.

با گرفتن یک نمونه ۲۰ تایی از جامعه ۲۰۵ تایی بازه اطمینان بصورت زیر خواهد بود و درنتیجه فرض ما رد نمی‌شود.

```
Confidence interval for length of the cars: (23.85,27.99)
The assumption is NOT FALSE
```

آزمون فرض دوم: میانگین ارتفاع ماشین ها ۵۵ است.

مشابه قسمت قبل این بار با گرفتن یک نمونه ۵۰ تایی از جامعه ۲۰۵ تایی بازه اطمینان بصورت زیر خواهد بود و درنتیجه فرض ما رد میشود.

```
Confidence interval for length of the cars: (52.64,53.96)
The assumption is FALSE
```

قسمت ج)

$$\varepsilon_i = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i-\theta x_i)^2}{2\sigma^2}} \rightarrow \prod_{i=1}^{n} \varepsilon_i = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i-\theta x_i)^2}{2\sigma^2}}\right)$$

$$\rightarrow \prod_{i=1}^{n} \varepsilon_i = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma}}\right) e^{-\sum_{i=1}^{n} \frac{(y_i-\theta x_i)^2}{2\sigma^2}}$$

$$\rightarrow ln \rightarrow \ln(f) = n.\ln\left(\frac{1}{\sqrt{2\pi\sigma}}\right).\frac{\left(-\sum_{i=1}^{n}(y_i-\theta x_i)^2\right)}{2\sigma^2}$$

$$MLE \rightarrow \frac{d}{d\theta} \ldots = 0 \rightarrow \sum_{i=1}^{n} x_i(y_i-\theta x_i) = 0 \rightarrow \sum_{i=1}^{n} x_i y_i - \theta \sum_{i=1}^{n} x_i^2 = 0$$

$$\rightarrow \hat{\theta}_{MLE} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$$

معادله خط رگرسیون نیز بصورت زیر است:

$$Y - \bar{Y} = \hat{\theta}(X - \bar{X})$$

$$Y - \frac{1}{n}\sum_{i=1} Y_i = \hat{\theta}\left(X - \frac{1}{n}\sum_{i=1} X_i\right)$$

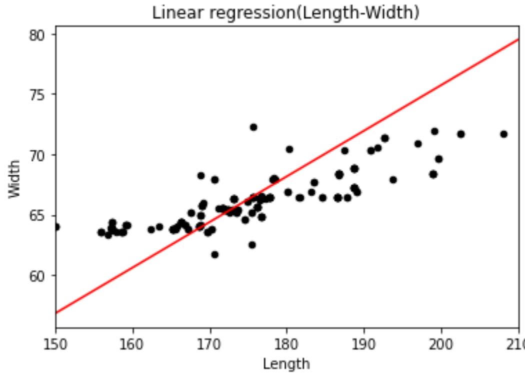
در نتیجه رابطه خطی بین دو متغیر بصورت زیر خواهد بود:

$$Y = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}\left(X - \frac{1}{n}\sum_{i=1} X_i\right) + \frac{1}{n}\sum_{i=1} Y_i$$

$$\rightarrow Y = \frac{cov(X,Y)}{var(X)} X\big(1 - mean(X)\big) + mean(Y)$$

حال به کمک اسکریپت پایتون بخش Part c قسمت ۶ ، دو نمودار را بررسی میکنیم و نتایج بصورت زیر خواهد بود:

۱. خط رگرسیون عرض برحسب طول ماشین:



۲. خط رگرسیون سوخت درون شهری برحسب برون شهری ماشین: