به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

آمار و احتمال مهندسی

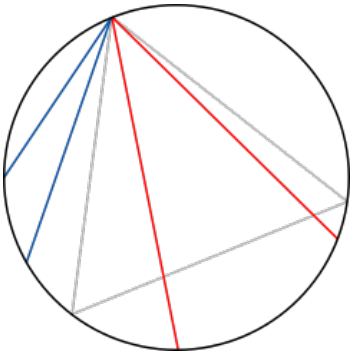
پروژه‌ی نهایی

طراحان:	روزبه نهاوندی، علیرضا جاوید، شیوا شاکری
تاریخ آپلود پروژه	۳۰ اردیبهشت ۱۴۰۱

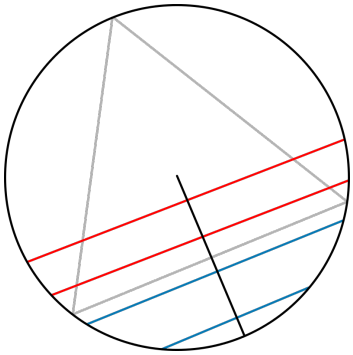
فهرست مطالب

۱ پارادوکس برتراند

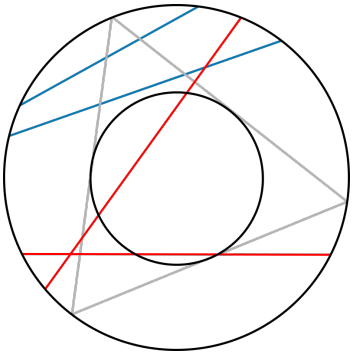
در برخی مسائل احتمالاتی، برای محاسبه ی احتمال یک پیشامد می توان از روش های متفاوت استفاده کرد که نتایج متفاوت و درستی را نیز به دنبال دارند؛ یکی از این مسائل را جوزف برتراند در سال ۱۸۸۹ مطرح کرد. مسأله ای که او بیان کرد را می توان با سه زاویه ی دید متفاوت حل کرد و به سه پاسخ متفاوت رسید. برای درک این پارادوکس، ابتدا مسأله را تعریف کرده و سپس به شبیه سازی راه حل های متفاوت آن مسائل دیگر می پردازیم. تعریف مساله. فرض کنید به طور تصادفی یک وتر را در دایره انتخاب کنیم. احتمال اینکه طول این وتر، بزرگتر از طول یک ضلع مثلث متساوی الاضلاع محاط در آن دایره باشد، چقدر است؟



(آ) شکل ۱ : راه حل اول



(ب) شکل ۲ : راه حل دوم



(ج) شکل ۳ : راه حل سوم

۱.۱ : راه حل اول باتوجه به تقارن، برای رسم یک وتر تصادفی، می توان ابتدا دو نقطه ی تصادفی روی محیط دایره انتخاب کرد و آن ها را به هم وصل کرد تا وتر بین این دو نقطه حاصل شود. نقطه اول را A و نقطه ی دوم را D در نظر بگیرید. فرض کنید A یکی از رئوس مثلث متساوی الاضلاع محاط باشد؛

۱. به صورت حل دستی احتمال اینکه وتر AD بزرگتر از طول ضلع مثلث ABC باشد را به دست آورید.

۲. برای شبیه سازی این راه حل، ابتدا ۱۰۰۰ بار، دو متغیر تصادفی یکنواخت در باز $[0, 2\pi]$ که معرف زاویه ی شعاع مربوط به دو نقطه ی تصادفی روی محیط یک دایره با شعاع واحد و مرکز مبدا هستند را تعریف کنید. (شکل ۴) مختصات نقاط انتهایی، ابتدایی و طول وتر آن را برحسب متغیرهای تصادفی بدست آورید . سپس با استفاده از کتابخانه ی Matplotlib دایره ی مورد نظر و وترهای تولید شده را رسم کنید. با بدست آوردن اندازه ی ضلع مثلث متساوی الاضلاع محاط در این دایره، نسبت تعداد وتر های بلندتر از یک ضلع مثلث به تعداد کل وترها را به دست آورید

۲.۱ راه حل دوم: باتوجه به تقارن، برای رسم یک وتر تصادفی، نقطه ای تصادفی روی محیط دایره انتخاب کرده و آن را به مرکز دایره وصل می کنیم. به این طریق توانسته ایم یک شعاع تصادفی از دایره انتخاب کنیم. سپس نقطه ای تصادفی از روی این شعاع انتخاب می کنیم. وتری وجود دارد که این شعاع در این نقطه، عمود منصف آن است.

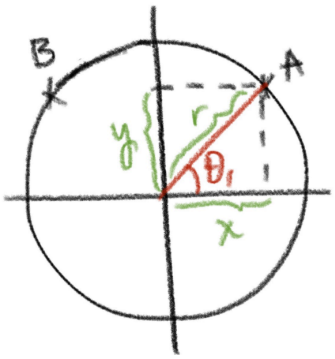
۱. به صورت حل دستی احتمال اینکه این وتر بزرگتر از طول ضلع مثلث ABC باشد را به دست آورید.

۲. برای شبیه سازی این قسمت نیز باید به تعداد ۱۰۰۰ بار، دو متغیر تصادفی یکنواخت تعریف کنید که یک متغیر تصادفی زاویه (،) و دیگری شعاع تصادفی $(P \in (0, r))$ ، که نسبتی از شعاع واحد است. باتوجه به شکل ۵ ،طول Q P را به دست آورید. سپس مختصات نقاط انتهایی، ابتدایی و طول وتر آن را برحسب متغیرهای تصادفی بدست آورید. سپس با استفاده از کتابخانه یatplotlib دایرهM ی مورد نظر و وترهای تولید شده را رسم کنید و نسبت تعداد وتر های بلندتر از یک ضلع مثلث به تعداد کل وتر ها را محاسبه کنید.

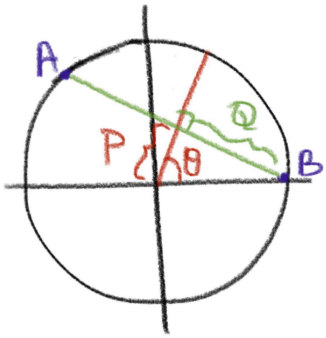
۳.۱ راه حل سوم: برای رسم یک وتر تصادفی، نقطه ای تصادفی داخل دایره انتخاب می کنیم و به مرکز دایره وصل می کنیم. سپس وتر عمود انتخابی را رسم می کنیم.

۱. به صورت حل دستی احتمال اینکه این وتر بزرگتر از طول ضلع مثلث ABC باشد را به دست آورید.

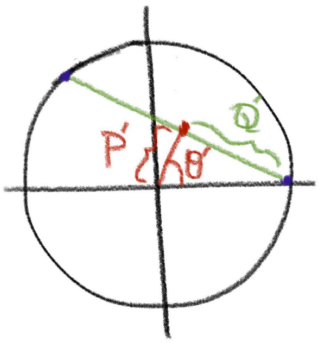
۲. برای شبیه سازی این قسمت، یک فرآیند پواسون همگن روی دایره انجام می شود تا یک نقطه را به صورت یکنواخت روی دایره انتخاب کنید. یک متغیر تصادفی (θ') مانند قبل در نظر بگیرید. برای تولید متغیر تصادفی دیگر (P')، یک متغیر تصادفی یکنواخت در بازه‌ی واحد ایجاد کنید و سپس جذر آن را در شعاع ضرب کنید. توضیح دهید چرا باید از متغیر تصادفی یکنواخت ایجاد شده جذر بگیرید؟ سپس مقادیر P', Q' را به دست آورید. سپس مختصات نقاط انتهایی، ابتدایی و طول وتر آن را برحسب متغیرهای تصادفی بدست آورید. حال با مقادیر استفاده از کتابخانه یMatplotlib دایره ی مورد نظر و وترهای تولید شده را رسم کنید و نسبت تعداد وتر های بلندتر از یک ضلع مثلث به تعداد کل وتر ها را محاسبه کنید.



(آ) شکل ۴ : شبیه‌سازی راه حل اول



(ب) شکل ۵ : شبیه‌سازی راه حل دوم



(ج) شکل ۶ : شبیه‌سازی راه حل سوم

۲ تخمین عدد π با روش مونت کارلو

روش های مونت کارلو به الگوریتم هایی گفته می شود که با استفاده از قوانین آمار و احتمال به حل مسائلی می پردازد که بصورت عادی حل آن ها دشوار است. در این مسئله می خواهیم عدد π را با این روش تخمین بزنیم.

فرض کنید در حال یک بازی دارت هستید و صفحه دارت بصورت یک مربع با ضلع ۱ متر می باشد.(همانگونه که در شکل ۲ مشخص شده است) امکان برخورد هر دارت به صفحه به صورت تصادفی و هم احتمال می باشد . اگر تعداد زیادی دارت پرتاب کنیم کسری از دارت ها که در ربع دایره مشخص شده قرار می گیرند به مقدار

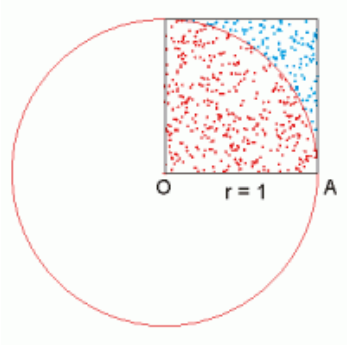
$$\frac{Area\ of\ quarter\ circle}{Area\ of\ square} = \frac{\pi}{4}$$

نزدیک می شود. بنابراین با محاسبه این کسر برای تعداد زیادی دارت می توانیم مقدار این کسر را محاسبه کرده و در نهایت تخمین مناسبی برای عدد π ارائه دهیم.

۱. برای $N = 100$ ، $N = 1000$ ، $N = 100000$ مانند شکل رسم شده توزیع دارت های پرتاب شده را نشان دهید. سعی کنید نقاطی که درون ربع دایره قرار می گیرند با دیگر نقاط متفاوت باشند.

۲. مقدار کسر را برای هر کدام از موارد الف محاسبه کرده و مقدار π را با استفاده از آن تخمین بزنید.

۳. سرعت همگرایی این روش را تا $N = 1000$ بررسی کنید و در یک نمودار آن را نشان دهید.



شکل ۳: تخمین عدد π

۴. تخمین ارائه شده در قسمت قبل تابعی از نمونه های گرفته شده از محیط است. در نتیجه این تخمین، خود یک متغیر تصادفی می باشد به عبارت دیگر اگر اسکریپت قسمت قبل را چندبار اجرا کنید،احتمالا مقادیرمتفاوتی دریافت خواهیدکرد. فرض کنید این متغیر تصادفی،توزیعی گاوسی داشته باشد. برای تخمین مقادیر امید ریاضی و واریانس آن می بایست چند نمونه از آن گرفت و سپس با استفاده از تخمین بیشینه درست نمایی ۲ مقادیر میانگین و واریانس را برای یک n ثابت و بزرگ محاسبه کرد. توزیع تخمینی از متغیر تخمین ارائه شده در قسمت قبل را به دست آورید.

۵. در نهایت می توان امید ریاضی توزیع تخمین زده شده را به عنوان تخمین بهتر ارائه کرد و واریانس تخمین گاوسی فوق معیاری از عدم قطعیت ما نسبت به تخمینی است که ارائه کرده ایم. بدیهیست که عدم قطعیت ما هیچ گاه صفر نخواهد شد، به عنوان مثال یکی از اولین عوامل ایجاد کننده عدم قطعیت تقریبیست که روی تعداد نمونه ها اعمال کردیم. می دانیم که برای دقیق بودن تعداد نمونه ها باید به بی نهایت میل کند ولی ما در بخش های قبل تعداد نمونه ها را محدود فرض کردیم. پس می توان انتظار داشت هرچه n را بزرگتر کنیم، عدم قطعیت از تخمین کمتر شود و به صفر میل کند. برای ارزیابی این فرضیه، تعداد نمونه ها را بین ۱ تا ۱۰۰۰ تغییر دهید و به ازای هر مقدار n، امید ریاضی و واریانس توزیع تصادفی تخمین را محاسبه کنید. سپس در یک نمودار، مقادیر امید ریاضی را برحسب تعداد نمونه ها رسم کنید

سپس به ازای هر داده روی نمودار، با یک خط عمودی بازه ی اطمینان ۹۵ درصدی داده ها را مشخص کنید.آیا فرضیه ای که مطرح کردیم تایید شد؟ اگر پاسخ منفیست، چرائی این موضوع را بررسی کنید.

۳ تخمین عدد اویلر

تا الان در بسیاری از در این بخش می خواهیم به سراغ تخمین یکی از پرکاربرد ترین اعدادی که تابحال در زندگی خود داشتید، برویم. احتمالا موارد با عدد اویلر e کلنجار رفته اید.این عدد ثابتی است که مقداری حدود ۲.۷۱۸۲۸ دارد. این ثابت اولین بار توسط ژاکوب برنولی در سال ۱۶۸۳ کشف شد. رابطه ای که در آن برنولی توانست این ثابت را پیدا کند، به شکل زیر بود.

$$\lim_{n\to\infty}(1+\frac{1}{n})^n$$

هماطور که قطعا می‌دانید، این عدد کاربرد های بسیار زیادی در ریاضیات، حساب دیفرانسیل و حتی توزیع های آماری که با آن ها آشنا هستید، دارد. حال به روش تخمینی که ما در نظر داریم، می پردازیم. شخصی که علاقه ی زیادی به توزیع یک نواخت دارد ادعا کرده است که اگر پشت سیستمی بشیند و بار ها توزیع یک نواخت در بازه $[0,1]$ ایجاد کند و اعداد ایجاد شده از این توزیع را با هم جمع کند تا بزرگ تر از یک شوند؛ به طور میانگین این شخص می بایس $e=2.71828$ بار توزیع یک نواخت ساخته تا مجموع این اعداد از یک بیش تر شوند. حال می خواهیم ببینیم که حدس این شخص درست هست یا خیر.

۱. فرض کنید یک توزیع یک نواخت بین ۰ و ۱ داریم و هربار، یک عدد تصادفی از این توزیع استخراج میکنیم و مقدار آن را با اعداد قبلی استخراج شده جمع میکنیم. اینکار را تا وقتی ادامه میدهیم که جمع این اعداد از بیشترk شود و سپس تعداد دفعاتی که از این توزیع عدد استخراج کردیم را داخل یک متغیر ذخیره میکنیم. میانگین دفعاتی که لازم است تا جمع اعداد استخراج شده از ۱ بیشتر شود را به صورت دستی بهدست آورید

۲. با استفاده از شبیه سازی، این میانگین را به دست آورید.

۳. به ازای $k=2,5,10,20,30,50$ این میانگین را به دست آورید و نمودار این میانگین‌ها را بر حسب k رسم کنید. حال با توجه به نمودار، یک فرمول برای تعداد دفعاتی که لازم است تا مجموع اعداد از k بزرگتر شود بیابید.

۴ تولید نمونه های تصادفی

در این تمرین روش مهمی را برای تولید نمونه های تصادفی یک توزیع با داشتن تابع pdf آن را بررسی می کنیم.

۱. در ابتدا نشان دهید اگر X یک متغیر تصادفی پیوسته با توزیع دلخواه باشد، بصورتی که تابع cdf آن برابر $F_X(x)$ باشد متغیر تصادفی

$$Y = F_X(x)$$

دارای توزیع یکنواخت در بازه (۰،۱) می باشد.

۲. با استفاده از قضیه بالا یک تابع random number generator بسازید به صورتی که نمونه های تصادفی با توزیع نمایی

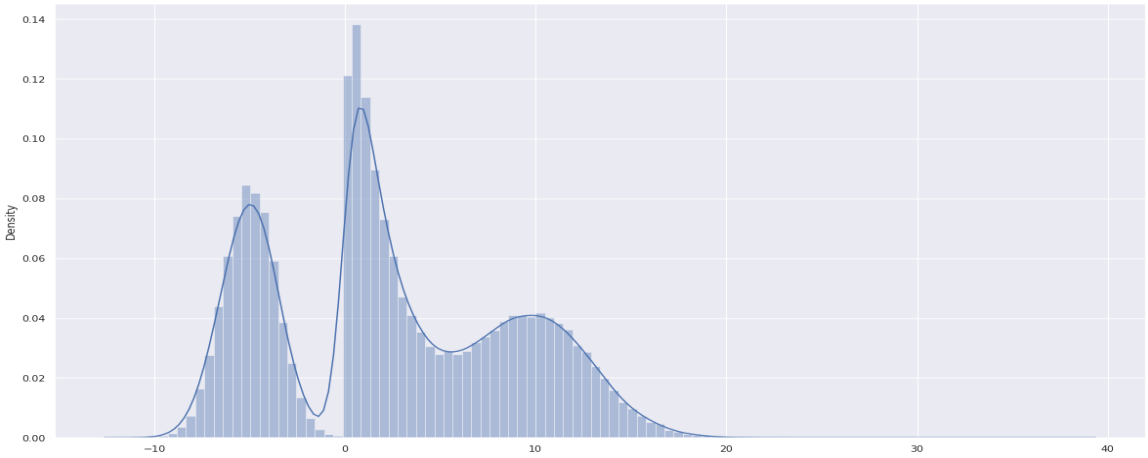
$$f_X(x) = \alpha e^{-\alpha x} u(x)$$

تولید کند. برای چک کردن درستی تابعی که طراحی کردید هیستوگرام نمونه های تولید شده را رسم و بررسی نمایید.(فرض کنید $\alpha = 0.15$ می باشد)

۳. با روش گفته شده در سوال، یک random number generator دیگر برای توزیع هندسی که یک توزیع گسسته است نیز طراحی کنید و مانند بخش قبل نمودار هیستوگرام آن را رسم کنید.(فرض کنید $p = 0.55$ می باشد)

۴. حال برای توزیع پیوسته غیر مرسوم زیر، تابع random number generator را ساخته و نمودار هیستوگرام آن را رسم کنید.

$$f_X(x) = \frac{3}{10} \times \frac{1}{\sqrt{2\pi \times 9}} e^{\frac{-(x-10)^2}{18}} + \frac{3}{10} \times \frac{1}{\sqrt{2\pi \times 2}} e^{\frac{-(x+5)^2}{4}} + \frac{4}{10} \times 0.45 e^{-0.45x} u(x)$$



شکل ۴: نمونه برداری

۵ SECRETARY PROBLEM

در این سوال می‌خواهیم به بررسی Secretary problem بپردازیم. در این مسئله فرض می‌شود که یک مدیر داریم که می‌خواهد از بین n کاندید، بهترین آنها را انتخاب کند. کاندید ها یکی پس از دیگری و به صورت تصادفی مصاحبه می‌شوند. تصمیم قبول یا رد شدن هر کاندید بلافاصله بعد از مصاحبه گرفته می‌شود. هنگامی که یک کاندید رد می‌شود، امکان قبولی دوباره آن شخص وجود ندارد. در حین مصاحبه، مدیر اطلاعات لازم برای طبقه بندی کاندیدهایی که تاکنون مصاحبه شده‌اند را به دست می‌آورد اما از کیفیت کاندید های بعدی ناآگاه است. هدف در این مسئله، یافتن بهرین استراتژی برای بیشینه کردن احتمال گزینش بهترین کاندید در بین n کاندید است. یک استراتژی، رد کردن k کاندید اول و سپس انتخاب اولین کاندیدی است که از این k کاندید رد شده بهتر باشد. اما باید توجه کرد که اگر k عددی کوچک باشد، اطلاعات کافی برای انتخاب بهترین کاندید را نداریم و اگر k بزرگ باشد، امکان اینکه بهترین کاندید در بین k کاندید رد شده باشد افزایش مییابد.

۱. به ازای n = ۱۰۰ یک لیست از اعداد ۰ تا n-۱ تولید کنید. این لیست نشان‌دهنده n کاندید است که کیفیت این کاندیدها در این لیست آورده شده‌اند. حال به ازای $k = range(5, 101, 5)$ و برای ۱۰۰۰۰ آزمایش، احتمال انتخاب بهترین کاندید را به ازای رد کردن k کاندید اول به دست آورید و در یک scatter plot رسم کنید، سپس بهترین عدد به‌دست آمده برای k را گزارش کنید.

۲. به ازای $k = n/e$ که e عدد نپر است و به ازای ۳ تا ۱۰۰ کاندید، احتمال انتخاب بهترین کاندید را بعد از رد کردن k کاندید اول در ۱۰۰۰۰ آزمایش به دست آورید و plot scatter این احتمالات را رسم کنید.

۳. به صورت دستی، احتمال انتخاب بهترین کاندید را به دست آورید.

۶ کار با داده

۱. در این پروژه، ما قصد داریم مجموعه‌ای از داده‌های واقعی را با آن‌چه در این درس آموخته‌اید بررسی و تحلیل کنیم. برای شروع تجزیه و تحلیل یک مجموعه داده، اولین قدم آشنایی با آن است. در اولین قدم می‌توان با مشاهده مواردی مثل ویژگی‌های مجموعه داده و توزیع مقادیر و تجسم داده‌ها برای حدس زدن اولیه در مورد آن آشنایی را انجام داد. در مرحله‌ی بعدی با انجام آزمایشات آماری، اطمینان حاصل می‌کنیم که حدس‌هایمان درست است و ادعاهای خود را با اطمینان بیان می‌کنیم.

برای این سوال از دیتا ست StudentPerformance استفاده کنید.

(آ) به طور خلاصه دیتاست داده‌شده را توصیف کنید و تعداد ویژگی‌های آن را بیان کنید.

(ب) با این توصیف، به نظر شما کدام ویژگی اطلاعات مهم‌تری دارد و چرا؟

یک ویژگی عددی از دیتاست انتخاب کنید و به سوالات زیر پاسخ دهید.

(آ) یک هیستوگرام با bin size مناسب بکشید و ویژگی‌های بارز توزیع آن را بیان کنید. از یک نمودار مناسب برای مقایسه کردن این توزیع با توزیع نرمال استفاده کنید.

(ب) چولگی یک متغیر تصادفی توصیف مناسبی از رابطه‌ی میانگین و میانه‌ی آن است. رابطه‌ی چولگی یک متغیر تصادفی را یافته و سعی کنید آن را توصیف کنید. حال چولگی را برای این متغیر انتخاب‌شده محاسبه کرده و نتیجه را توصیف کنید.

(ج) مقادیر میانگین، میانه، واریانس و انحراف معیار را محاسبه کرده و آن‌ها را توصیف کنید.

(د) نمودار density-plot را برای این متغیر بکشید و خطوط مربوط به میانگین و میانه را به آن اضافه کنید، حال ارتباط این خطوط را با نمودار رسم‌شده بیان کنید.

(ه) بازه اطمینان ۹۵ درصدی را برای میانگین این متغیر محاسبه کنید.

(و) هیستوگرام این متغیر را کشیده و میانگین هر نمونه را به صورت خطی عمودی در بالای هیستوگرام نمایش دهید. هم‌چنین بازه اطمینان را به صورت دو خط عمودی نشان دهید.

حال دو ویژگی عددی متفاوت از این دیتاست را انتخاب کرده و به سوالات زیر پاسخ دهید.

(آ) ابتدا بدون شبیه‌سازی و بر اساس فرضیات خود، ارتباط این دو ویژگی را از روی توضیح آن‌ها حدس زده و حدس خود را بیان کنید.

(ب) نمودار پراکندگی را برای این دو متغیر رسم کرده و ارتباط آن‌ها را با هم بیان کنید.

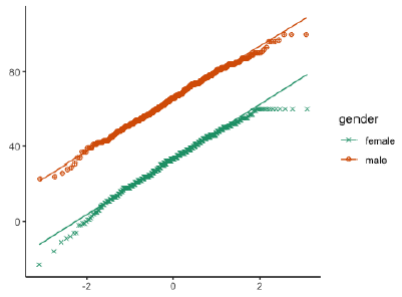
(ج) ضریب هم‌بستگی بین این دو متغیر را محاسبه کنید.

(د) حال حدسی که در قسمت آ زده بودید را با ضریب هم‌بستگی و ارتباط متغیرها مقایسه کنید.

۲. در این سوال قصد داریم با انواع نمودارهای پرکاربرد آشنا شویم. از دیتاست grades برای این کار استفاده می‌کنیم. این مجموعه، شامل نمرات دروس ریاضی، writing و reading برای ۱۰۰۰ دانش‌آموز است. توجه داشته باشید که برای کشیدن نمودارها می‌بایست از کتاب‌خانه‌ی ggplot استفاده نمایید. هم‌چنین نمودارهای شما باید همگی دارای برچسب‌های عنوان و محورهای مربوطه باشد.

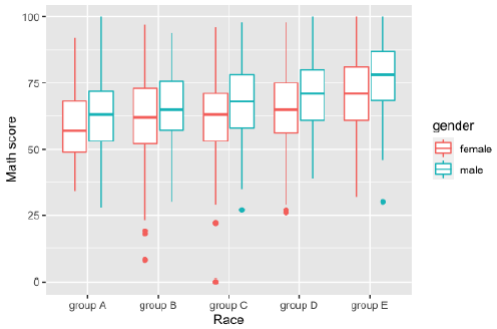
(آ) مجموعه داده را در ابتدا بخوانید، حال برای هر جنسیت، متغیر writing score را در قالب هیستوگرام نمایش دهید.

(ب) توزیع هر سه متغیر reading score ، writing score و math score را با کشیدن QQ-Plot آن‌ها بررسی کنید. آیا می‌توان این برداشت را کرد که این متغیرها توزیعی مشابه با توزیع گوسی دارند؟ توضیح دهید.
شکل شما باید شبیه شکل زیر باشد.



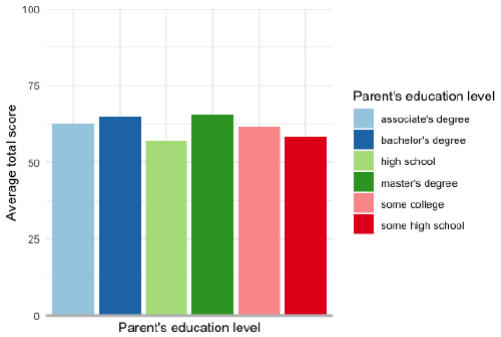
(ج) متغیر math score را در نظر بگیرید. حال برای همه‌ی دانش‌آموزان و جنسیت‌ها و گروه‌های مختلف، به طور مجزا boxplot های آن‌ها را بکشید.

شکل شما باید شبیه شکل زیر باشد.



(د) نمره‌ی نهایی (میانگین تمامی نمره‌ها) را برای هر دانش‌آموز محاسبه کرده و در یک ستون ذخیره کنید و آن ستون را به مجموعه‌ی داده اضافه کنید. حال یک bar plot رسم کرده به طوری که هر مستطیل در آن نشان‌دهنده‌ی میانگین نمره دانش‌آموزانی باشد که والدین آن‌ها دارای مدرک تحصیلی یکسانی هستند.

شکل شما باید شبیه شکل زیر باشد.



۷ توضیحات

۱. زبان برنامه نویسی خواسته شده در سوالات پایتون است.

۲. شما می بایست علاوه بر کدهای پیاده سازی شده، گزارشی تحلیلی و ریاضی از نتایج خود ارائه دهید. توجه داشته باشید که مفهوم گزارش پروژه با مفهوم توضیح کد متفاوت است در نتیجه در فایل گزارش، از درج کد جدا بپرهیزید.

۳. گزارش کار، اولین و مهم ترین آیتم نمره دهی می باشد در نتیجه با صرف زمان مناسب، گزارشی تهیه کنید که بازتاب گر زحماتی باشد که برای انجام پروژه کشیده اید. استفاده ی صحیح از نیم فاصله، علائم نگارشی، گویا بودن جملات و پاراگراف بندی مناسب از جمله مواردیست که در نگاه اول جلب توجه می کند و نکاتی نظیر استفاده از زیرنویس برای تصاویر و بالانویس برای جداول، ارجاع دادن به روابط و تصاویر با شماره ی مربوط به هر کدام و ... از جمله خصوصیت های یک نوشته ی آکادمیک است.

۴. در گزارش خود متون فارسی را با استفاده از فونت Nazanin B با سایز ۱۴ بنویسید.

۵. کدهای پایتون خود را حتما در قالب دفترچه‌ی ژوپیتر بارگذاری کنید و در نهایت یک فایل گزارش پی‌دی‌اف را در کنار دفترچه‌های ژوپیتر در قالب زیپ با نام "SID-FullName.zip" در صفحه‌ی درس بارگذاری کنید.

۶. ابهامات خود در مورد سوالات و یا قالب گزارش در گروه تلگرامی درس مطرح کنید. در انتهای هر پیام طراحان را منشن کنید. سوالات در گروه پرسیده شده و همان جا پاسخ داده خواهند شد تا در دسترس همه ی دانشجوها قرار بگیرند.

(آ) $@roozbeh\_n99$

(ب) $@alireza\_javid01$

(ج) $@shiivashakerii$

۷. دقت کنید تاریخ ددلاین پروژه ۳۱ خرداد روز سه‌شنبه است.