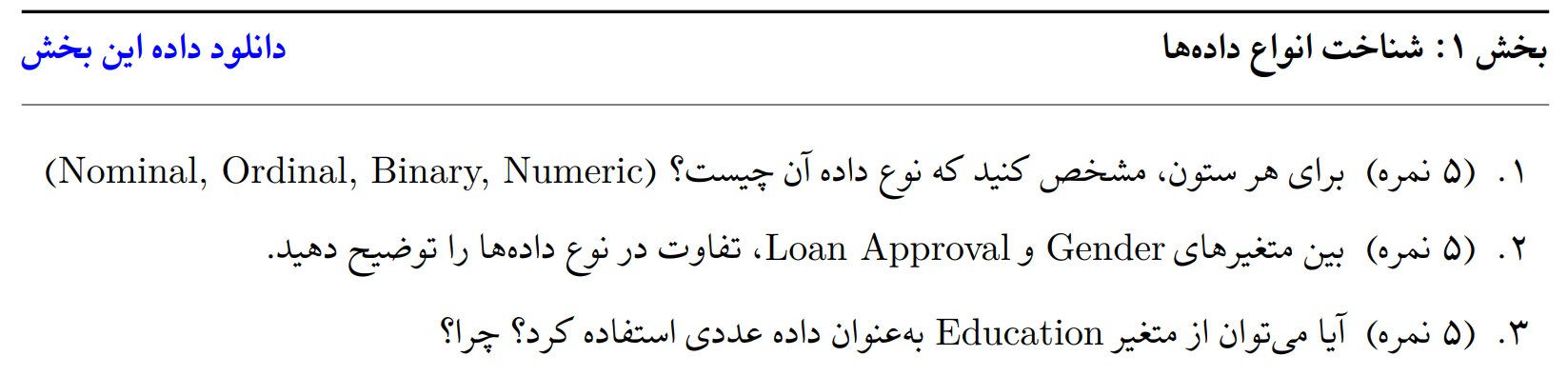
**گزارش تمرین اول مبانی داده‌کاوی و کاربردهای آن (۲۱۰۱۹)**

استاد: دکتر مانا مس‌کار  
دستیار آموزشی: محمدمهدی منتظری هدش  
اعضای گروه: صبا عبدی (۴۰۱۱۰۴۲۷6)، آوا صدیقی (۴۰۱۱۱۰۱۵۹)



|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CustomerID | Name | Age | Salary | Gender | Marital Status | Education | Loan Approval | Zip Code | Satisfaction Score |
| Nominal | Nominal | Numeric | Numeric | Binary | Nominal | Ordinal | Binary | Nominal | Ordinal |

CustomerID: در این قسمت از اعداد استفاده شده است. اما نکته مهم این است که ترتیب یا معنای عددی ندارند و شناسه‌ای برای شناسایی مشتریان است. بنابراین این attribute از نوع Nominal است.

Name: این قسمت نام مشتریان را شامل می‌شود و به همین علت از نوع Nominal است.

Age: این قسمت سن مشتریان را شامل می‌شود و به همین علت از نوع Numeric است.

Salary: این قسمت حقوق مشتریان را شامل می‌شود و به همین علت از نوع Numeric است.

Gender: این قسمت، جنسیت مشتریان را نشان می‌دهد. در این ستون تنها دو مقدار male و female داریم. به همین علت از نوع binary است.

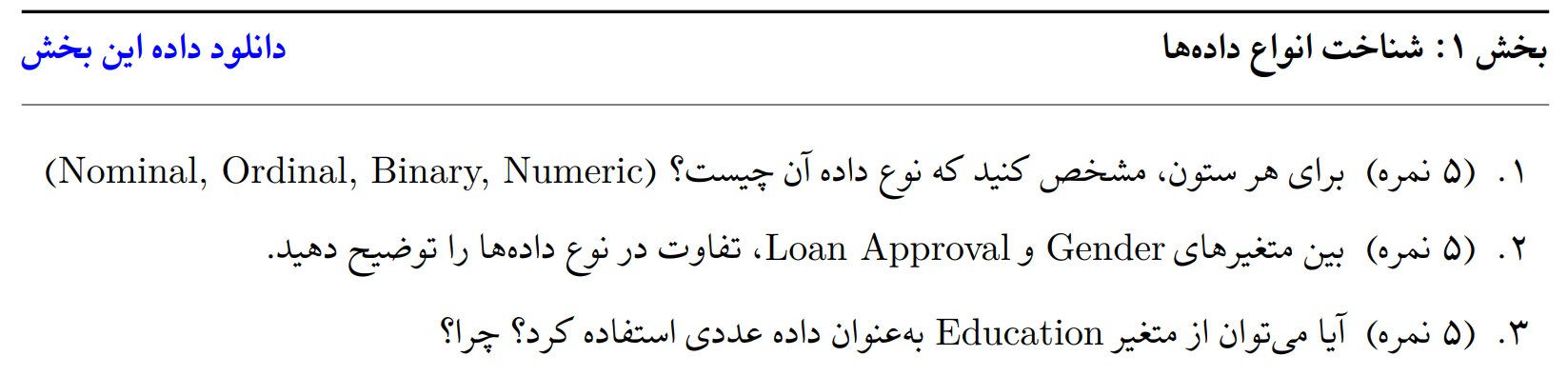
Marita Status: این قسمت وضعیت تأهل مشتریان را نشان می‌دهد و همچنین ترتیب خاصی ندارد. بنابراین از نوع Nominal است.

Education: این قسمت وضعیت تحصیلی مشتریان را نشان می‌دهد. به دلیل اینکه سطح تحصیلات به ترتیبی دنباله‌روی یکدیگر هستند، مثال پس از کارشناسی می‌توان کارشناسی ارشد را خواند، این ستون از نوع Ordinal است.

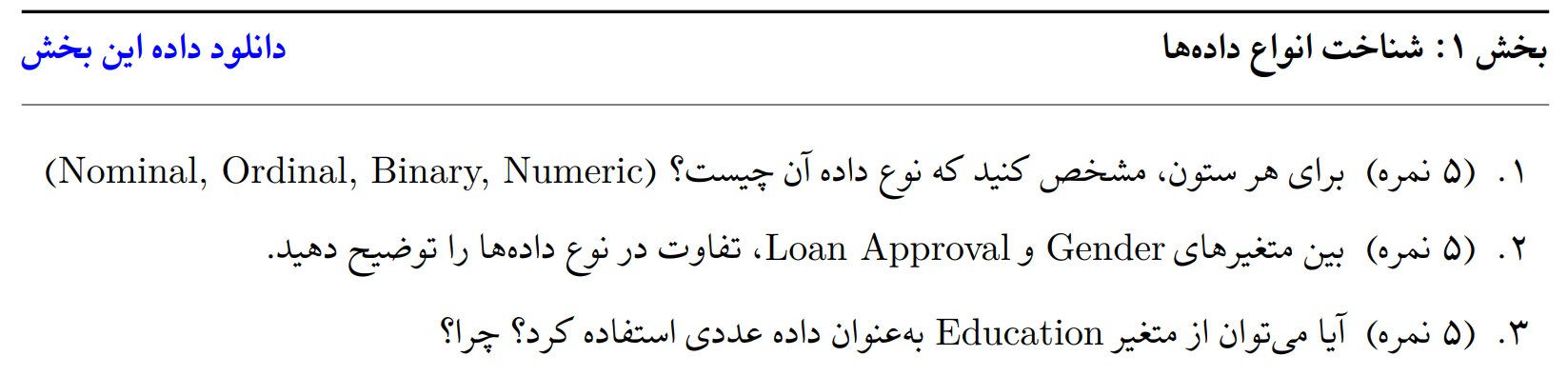
Loan Approval: این قسمت وضعیت وام درخواستی را نشان می‌دهد که تأیید یا رد شده است. به دلیل داشتن دو گزینه yes/no، این ستون از نوع binary است.

Zip Code: این قسمت، کدپستی مشتریان را نشان می‌دهد. کدپستی از اعداد تشکیل شده است اما به علت آنکه موقعیت مکانی را نشان داده و بار عددی ریاضی ندارد، آن را از نوع Nominal به شمار می‌آوریم.

Satisfaction Score: این قسمت میزان رضایت را نشان می‌دهد. به دلیل اینکه میزان رضایت از 1 تا 5 است و در دل خود ترتیبی جا داده است، این ستون را از نوع Ordinal به شمار می‌آوریم.

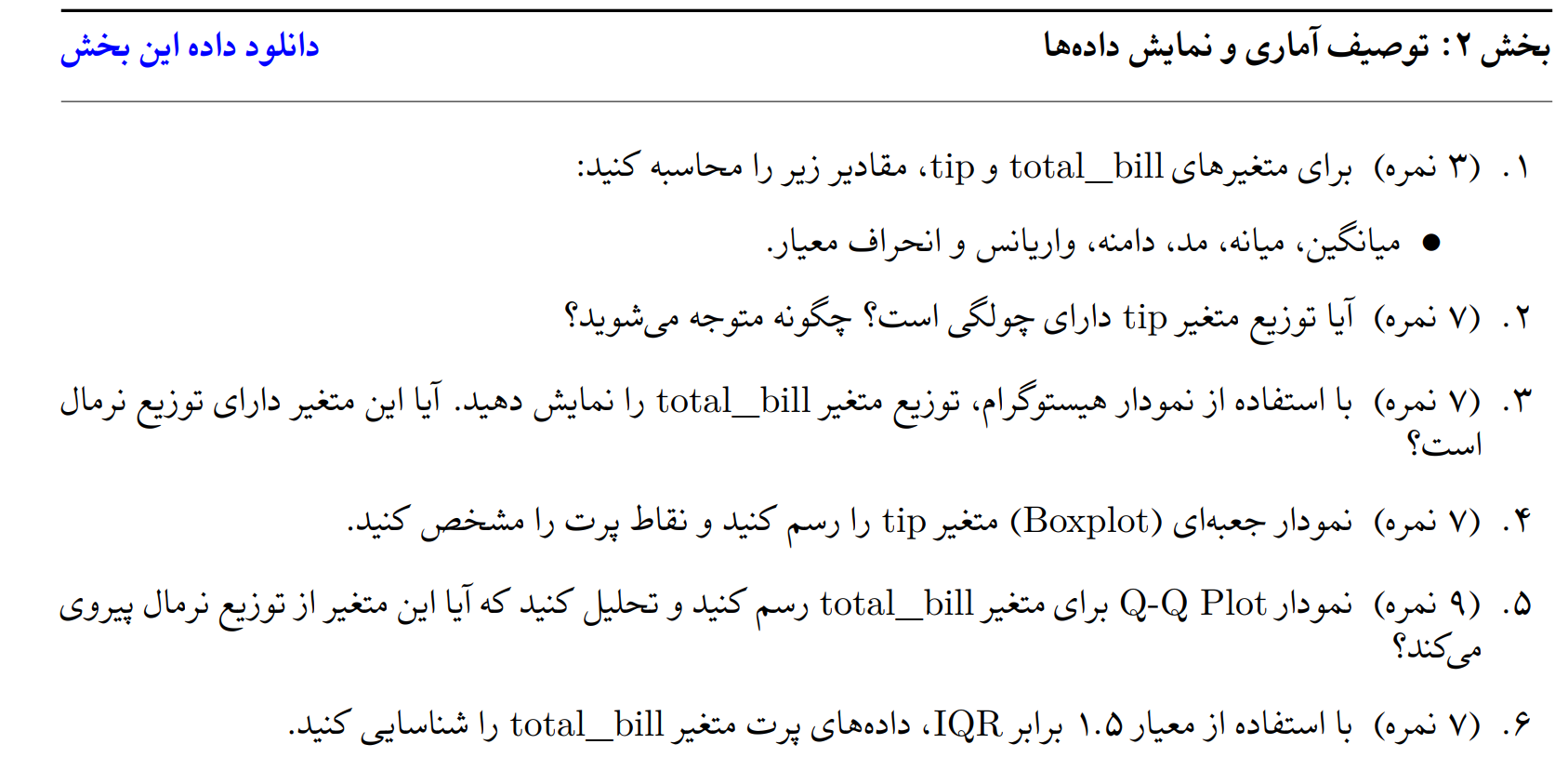


متغیرهای Gender و Loan Approval هر دو از نوع binary به شمار می‌آیند اما از برخی جنبه‌ها، با یکدیگر تفاوت‌هایی دارند که به شرح آن‌ها می‌پردازیم. در ابتدا، وام می‌تواند قبول یا رد شود؛ یعنی برای ستون Loan Approval نمی‌توان حالت دیگری را تصور کرد. اما در بحث Gender می‌توان موارد متعددی را در نظر گرفت و به نوعی آن را از حالت Binary خارج کرد و به Nominal تبدیل کرد یا رویکرد دودویی را به ابعاد بالاتر برد؛ به عنوان مثال، اگر Non-binary اضافه شود، آن را با 100، Male را با 010 و Female را 001 نشان داد که همین رویکرد نیز مشکلاتی را با خود ایجاد خواهد کرد. در ادامه، ستون Gender یک ویژگی توضیحی است و یک متغیر مستقل در تحلیل‌ها به شمار می‌رود. اما در Loan Approval نتیجه یک تصمیم اعلام می‌شود و یک متغیر وابسته است که متناسب با ورودی‌ها نتیجه آن مشخص می‌شود. علاوه بر این، attribute جنسیت یک متغیر binary متفارن است. یعنی صفر و یک آن بار تحلیلی خاصی را برای ما به ارمغان نمی‌آورد. این در حالی است که در Loan Approval، متناسب با تحلیل مدنظر، به هر یک از موارد yes و no مقدار یک را تخصیص داده و تحلیل خود را انجام می‌دهیمو

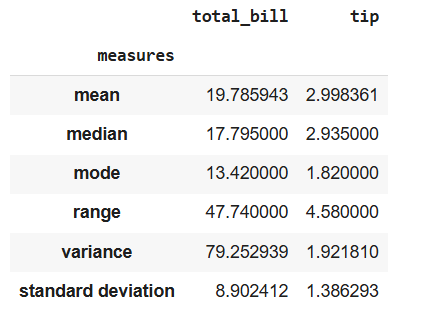


در سوال 1 این بخش، نوع ستون Education را مشخص کردیم که Ordinal بود. این ترتیب سطوح مختلف تحصیلات را نشان می‌دهند ولی نمی‌توان آن‌ها را به داده عددی تبدیل و استفاده کرد. در ابتدا باید بدانیم که فاصله بین مقاطع مختلف تحصیل با یکدیگر یکسان نیست. یعنی لزوماً مقدار عددی Phd 4 برابر مقطع High School نیست. بنابراین نمی‌توان آن‌ها را به عدد درستی تبدیل کرد. علاوه بر این، اگر فواصل مساوی را نیز بپذیریم، تحلیل‌های آماری بر روی این داده‌ها بی‌معنا خواهد بود. به عنوان مثال ما نمی‌دانیم میانگین 3.25 به چه معنی است و چگونه باید آن را تفسیر کنیم. بنابراین امکان تبدیل این مقادیر به داده عددی نیست.

لازم به ذکر است که می‌توان متناسب با رتبه‌بندی موجود، به مقاطع تحصیلی عددی تخصیص دهیم که این عدد رنک مقاطع را نشان دهد و غیر از این کاربرد، تبدیل دیگری نمی‌توان داشت.



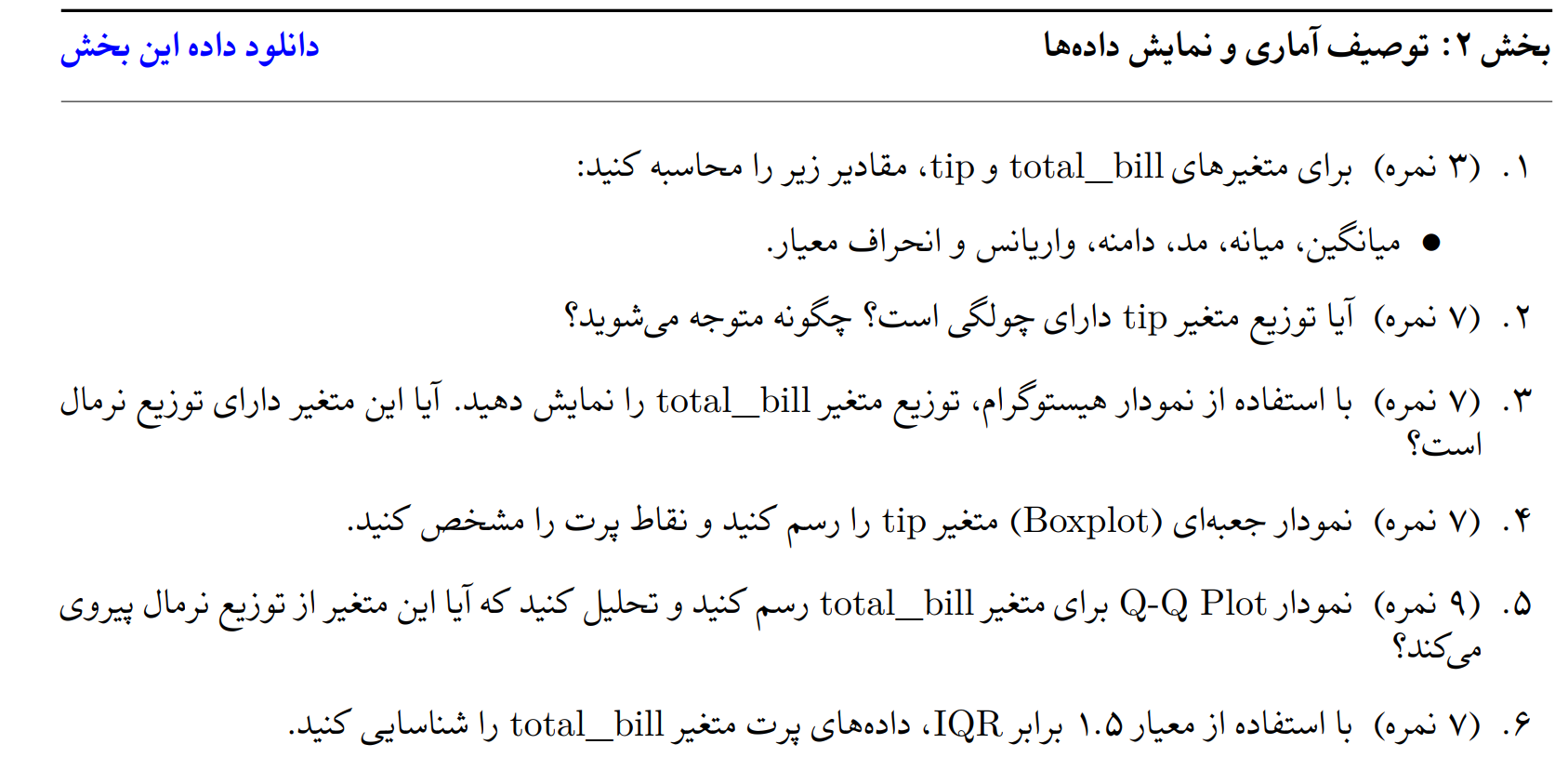
برای این قسمت از سوال از توابع کتابخانه pandas استفاده می‌کنیم. با استفاده از توابع اولیه چون mean، var و ... موارد خواسته شده را حساب می‌کنیم. حال اعداد به دست آمده برای هر ستون را تحلیل می‌کنیم.



ابتدا با ستون total\_bill آغاز می‌کنیم. میانگین این ستون برابر با 19.79، میانه این ستون برابر با 17.8 و مد این ستون برابر 13.42 است. مقادیر این سه معیار، از منظر مقدار و به ترتیب کاهشی، به شکل میانگین و میانه و مد هستند. میانگین نسبت به میانه کمی بیشتر است. به همین دلیل احتمال دارد که چند داده پرت با مقادیر بالا در این ستون وجود داشته باشند. همچنین به دلیل کمتر بودن مد از دو معیار دیگر، می‌توان گفت که این ستون چولگی مثبت دارد و داده‌ها در مقادیر کمتر تمرکز دارند. دامنه این ستون برابر با 47.74 است که نشان‌دهنده وجود پراکندگی زیاد در داده‌ها است. در نهایت، واریانس این ستون برابر 79.25 و انحراف معیار این ستون برابر 8.9 است. این اعداد، همانند دامنه، حاکی از پراکندگی زیادی در داده‌ها هستند؛ به نحوی که به طور میانگین داده‌ها چیزی در حدود 9 دلار از میانگین فاصله دارند.

در نهایت با ستون tips ادامه می‌دهیم. میانگین این ستون برابر با 3، میانه این ستون برابر با 2.94 و مد این ستون برابر با 1.82 است. در این اعداد تفاوت فاحشی، مثل ستون total\_bill به چشم نمی‌خورد. میانه به میانگین نزدیک است که این نشان‌دهنده یک‌دست بودن داده، تا حدودی، است. همچنین مد نیز کمی از میانگین کمتر است. دامنه این ستون برابر با 4.58، واریانس این ستون برابر 1.92 و انحراف معیار این ستون برابر با 1.39 است. تمامی این اعداد حاکی از این هستند که داده‌های موجود در این ستون پراکندگی زیاد و فاحشی نداشته و در اطراف میانگین به حالت پایداری قرار گرفته‌اند.

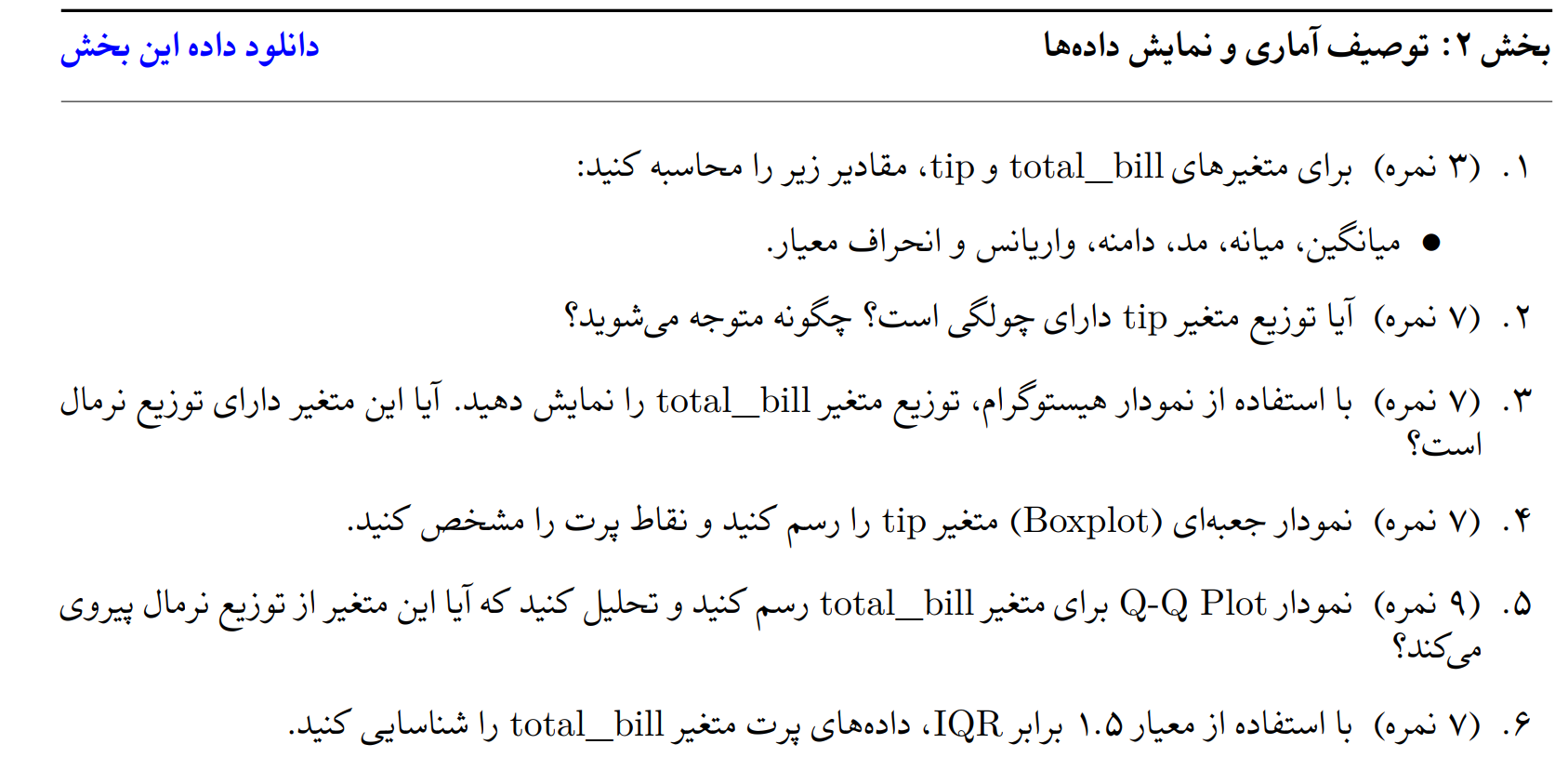
در نهایت می‌توان نتیجه گرفت که در ستون total\_bill شاهد پراکندگی زیاد و چولگی مثبت هستیم و در ستون tips داده‌ها مورد خاص و عجیبی ندارند.



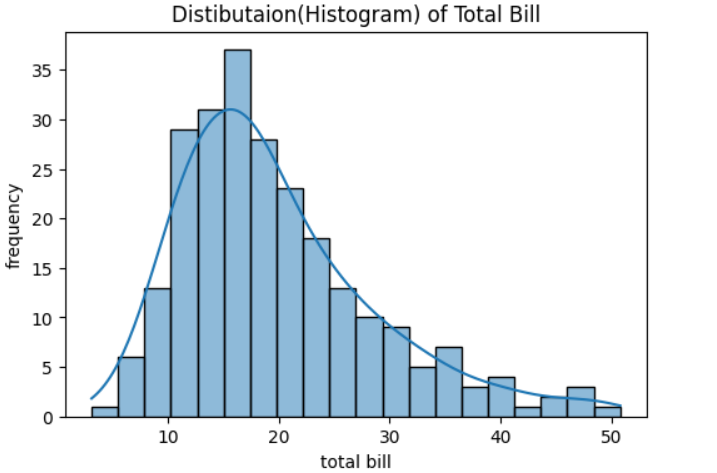
می‌دانیم چولگی زمانی اتفاق افتاده است که یکی از روابط میانگین< میانه < مد یا مد< میانه < میانگین برقرار باشد. در ستون tip حالت دوم، یعنی مد< میانه < میانگین، رخ داده است (به عکس جدول در سوال اول این بخش رجوع کنید.) . به دلیل اینکه این مقادیر به یکدیگر نزدیک هستند، از آزمون skewtest برای اطمینان بیشتر استفاده می‌کنیم. آماره آزمون برابر 0.29 است که از مقدار 0.05 بزرگتر است و فرض صفر، چولگی مثبت داشتن، را رد می‌کند.



متناسب با دو نکته‌ای که ذکر شد، بهترین نتیجه این است که بگوییم در این ستون چولگی مثبت وجود دارد ولی بسیار ضعیف است.

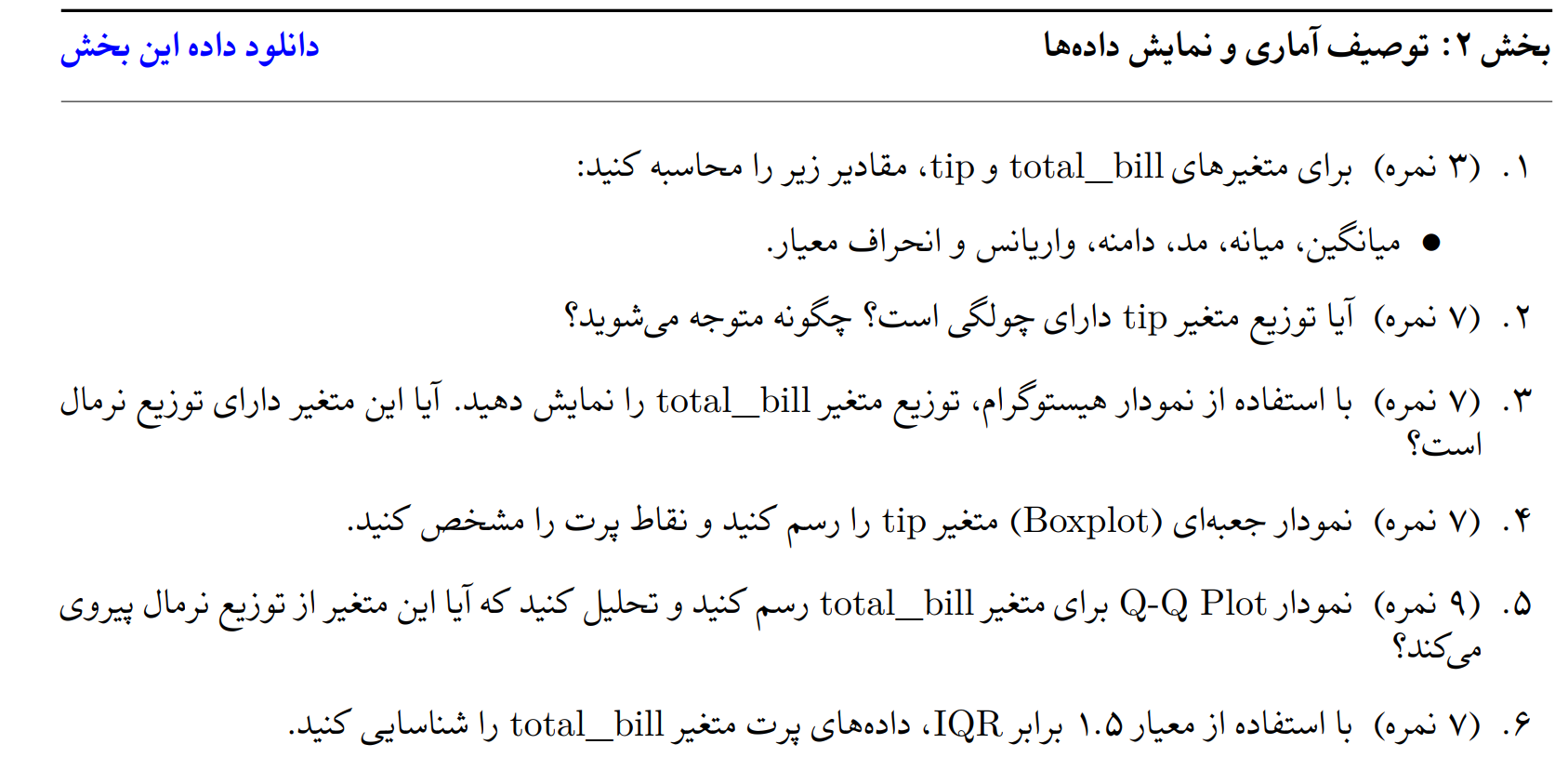


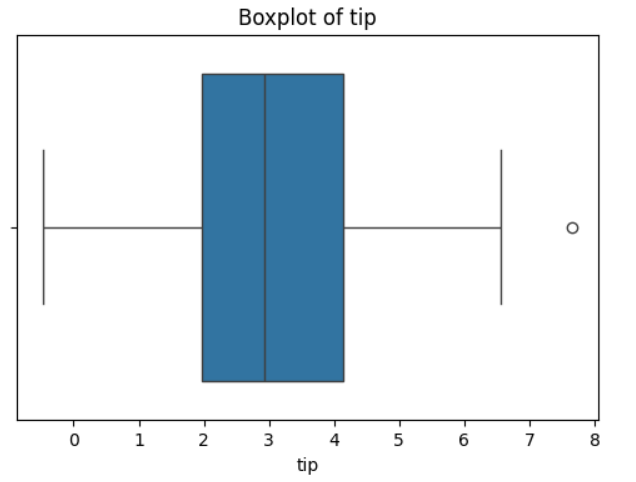
ابتدا با استفاده از کتابخانه‌های موجود، نمودار هیستوگرام را رسم می‌کنیم. از نمودار رسم شده می‌توان دریافت که از توزیع نرمال پیروی نمی‌کند.



برای اطمینان و اعتبار بیشتر تحلیل خود از آزمون Shapiro استفاده می‌کنیم. این آمون برای تعیین نرمال بودن یا نبودن داده‌های کوچک تا متوسط، همانند دیتاست ما، استفاده می‌شود. p value این آزمون برابر با 3.33 است که از مقدار 0.05 بسیار بزرگتر است. بنابراین با قطعیت می‌توان گفت که این داده‌های این ستون از توزیع نرمال پیروی نمی‌کنند.



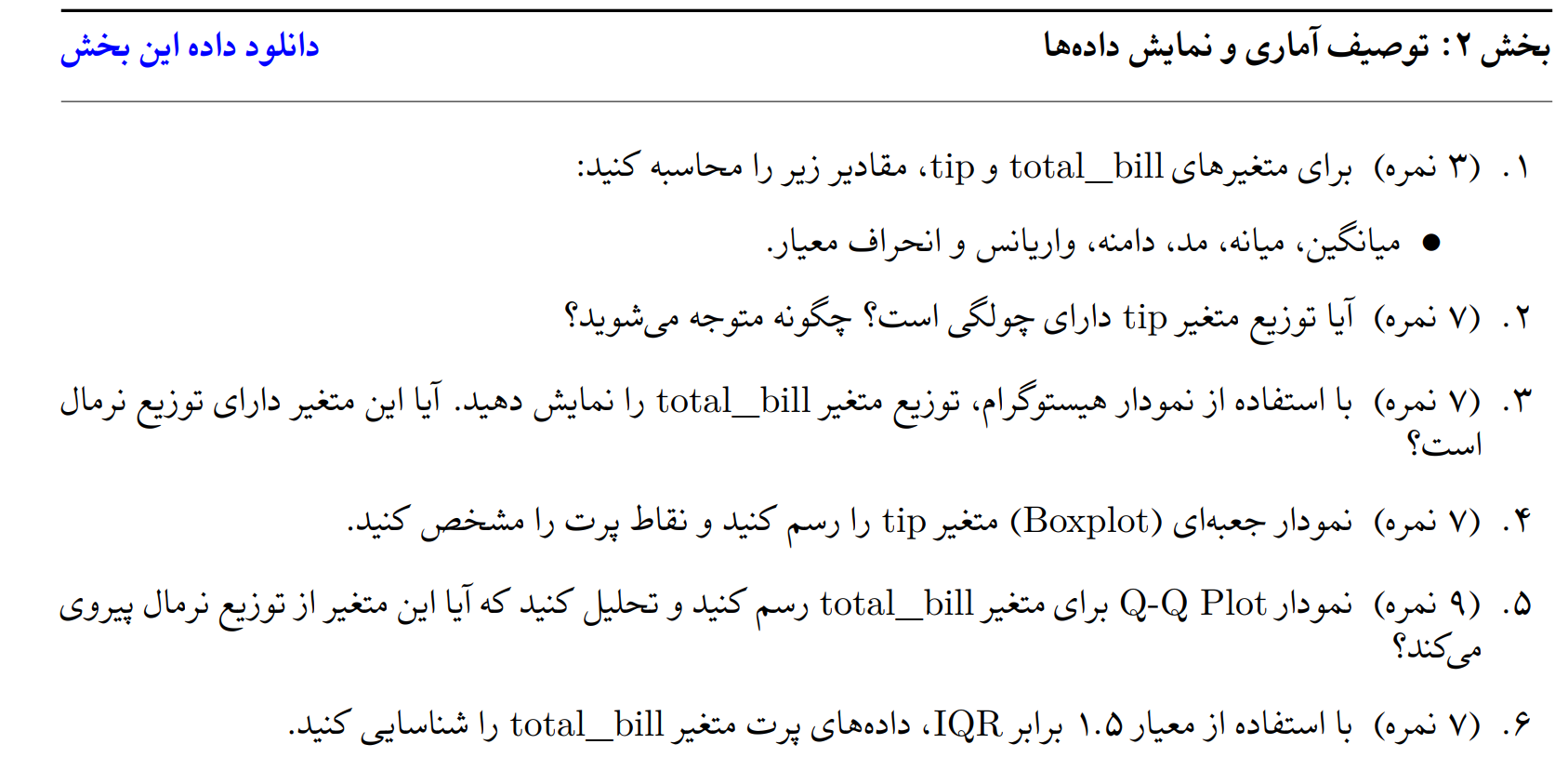
با استفاده از کتابخانه‌های مربوطه، نمودار جعبه‌ای ستون tip را رسم می‌کنیم.

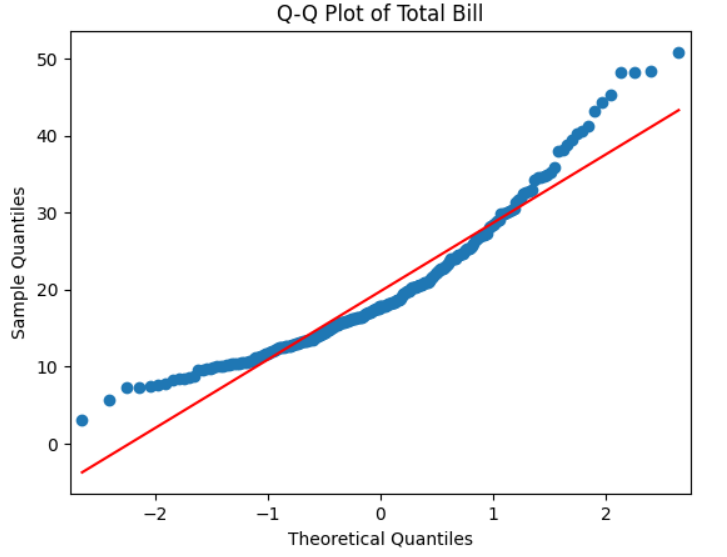


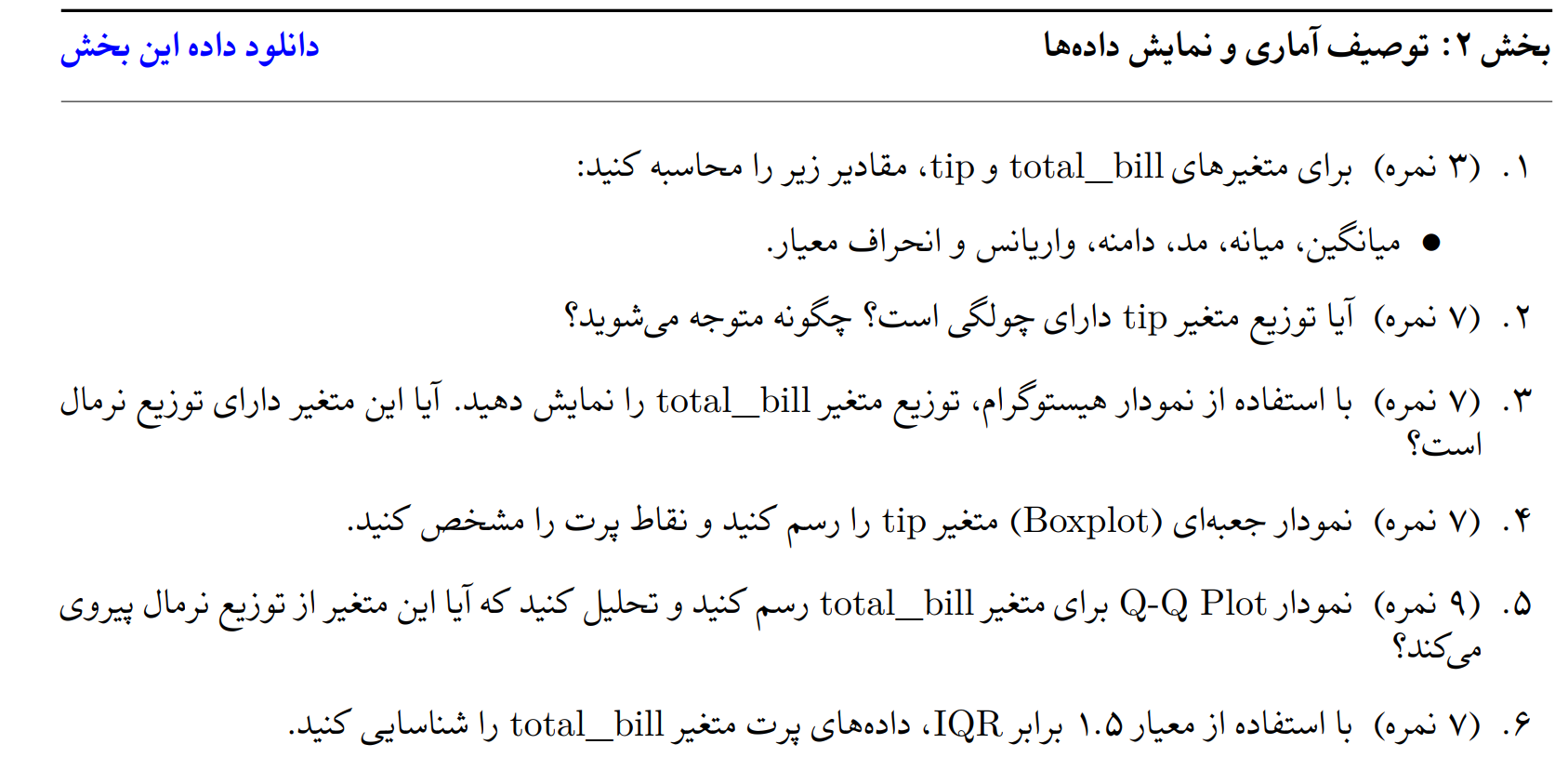
در شکل به دست آمده حدود بالا و پایین، چارک‌های اول و سوم و میانه مشخص هستند. طول جعبه نشان‌دهنده گستره میانی داده‌ها است و می‌توان بیان کرد که حجم خوبی از داده‌ها در این گستره قرار دارند. علاوه بر این چون میانه دقیقا وسط جعبه نیست، داده‌های ما تقارن کامل ندارند. به دلیل وجود یک داده پرت در سمت راست، می‌توان گفت که داده‌ها تا حدودی چولگی مثبت ضعیفی دارند. این چولگی ضعیف را کمتر بودن میانه نسبت به میانگین نیز تأیید می‌کند (دقت کنید که این تفاوت زیاد نیست؛ بنابراین چولگی بسیار ضعیف است اگر وجود داشته باشد؛ همانگونه که در بخش‌های بالاتر توضیح داده شد.).

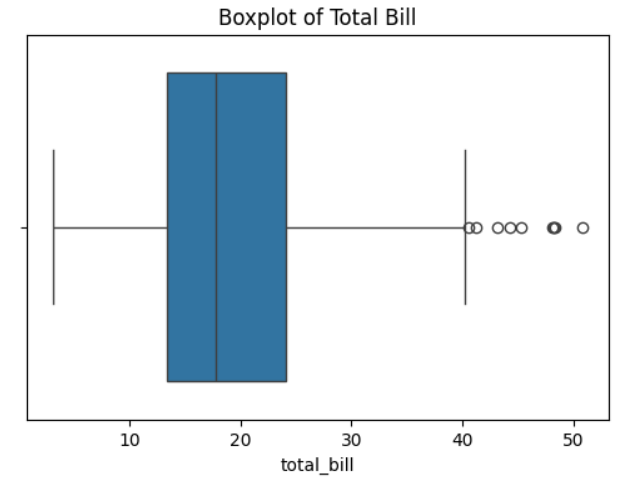
در شکل مشهود است که تنها یک نقطه از حدود مشخص خارج است؛ این نقطه داده پرت ما محسوب می‌شود. در سلول بعدی کد با محاسبه IQR و حدود نمودار جعبه‌ای، داده پرت موردنظر را اسلایس می‌کنیم. این داده، 210امین داده ( ایندکس‌گذاری از صفر است.) است. مقدار tip در این سطر 7.65 واحد است و از میانگین، میانه و مد فاصله زیادی داشته و حتی از دامنه به دست‌آمده برای این ستون بیشتر است.



برای رسم Q-Q Plot موردنظر از کتابخانه statmodel استفاده می‌کنیم. لازم به ذکر است که مولفه line=’s’ باعث می‌شود که داده دیگر مورد استفاده در این نمودار از توزیع نرمال باشد. حال نمودار را تحلیل می‌کنیم. در ابتدا، نقاط رفته رفته به خط نرمال نزدیک می‌شوند اما در ادامه از آن فاصله گرفته، دوباره به آن نزدیک شده و در انتها مجدد به سمت چپ و دور از خط نرمال تمایل پیدا می‌کنند. این روند نشان‌دهنده این است که ستون total\_bill از توزیع نرمال پیروی نمی‌کند. علاوه بر این متناسب با روند داده‌ها می‌توان بیان داشت که داده‌ها چولگی مثبت نیز دارند.



ابتدا برای درک بهتر این ستون نمودار جعبه‌ای آن را رسم می‌کنیم.



در نمودار به دست آمده تعدادی داده پرت در سمت راست نمودار جعبه‌ای قابل مشاهده است. در سلول بعدی کد با استفاده از معیار ذکر شده در صورت سوال این داده‌های پرت را شناسایی می‌کنیم. 9 داده به دست می‌آوریم که همگی از میانگین، میانه و دامنه بزرگتر هستند. لازم به ذکر است که این داده‌های پرت گواه بر چولگی مثبت این ستون هستند.

