A picture containing logo

Description automatically generated

محمد امین حاجی حسینی - 98724205



تمرین شماره 1 داده‌کاوی

فهرست مطالب

[نصب و راه‌اندازی 1](#_Toc55042381)

[دید‌‌کلی داده‌ها 2](#_Toc55042382)

[بررسی موارد جدید و کشته‌شدگان ایران در بازه‌های زمانی مختلف 3](#_Toc55042383)

[بررسی داده‌های گم‌شده 5](#_Toc55042384)

[بررسی نمودار‌ جعبه‌ای ایران و چهار کشور انتخابی 6](#_Toc55042385)

[بررسی مقادیر محاسبه‌شده برای کشور ایران 7](#_Toc55042386)

[بررسی داده‌های پرت 7](#_Toc55042387)

[بررسی نمودار جعبه‌ای کشور ایران برای موارد هفتگی جدید ابتلا 8](#_Toc55042388)

فهرست شکل‌ها

[شکل 1 - نصب کتابخانه‌های مورد نیاز 1](#_Toc55042389)

[شکل 2- اجرای کد 1](#_Toc55042390)

[شکل 3 - بخشی‌ای از خروجی کد 1](#_Toc55042391)

[شکل 4 - نمودار میله‌ای موارد و کشته‌شدگان جدید در ایران در به صورت ماهانه 3](#_Toc55042392)

[شکل 5 - نمودار میله‌ای موارد و کشته‌شدگان جدید در ایران در به هفتگی 4](#_Toc55042393)

[شکل 6 - نمودار جعبه‌ای کشور‌های ایران، افغانستان، عراق، انگلستان، ایتالیا 6](#_Toc55042394)

فهرست جداول

[جدول 1- بررسی 10 فیلد از مجموعه‌داده 2](#_Toc55042395)

[جدول 2- مقادیر محاسبه‌شده برای کشور ایران 7](#_Toc55042396)

[جدول 3 - 10 مورد از داده‌های پرت 7](#_Toc55042397)

# نصب و راه‌اندازی

در صورت تمایل، پس از نصب و راه‌اندازی یک محیط مجازی[[1]](#footnote-1) و فعال‌سازی آن، با استفاده از دستور موجود در شکل 1، نیازمندی‌های لازم جهت اجرای کد، به صورت خودکار نصب خواهند شد.



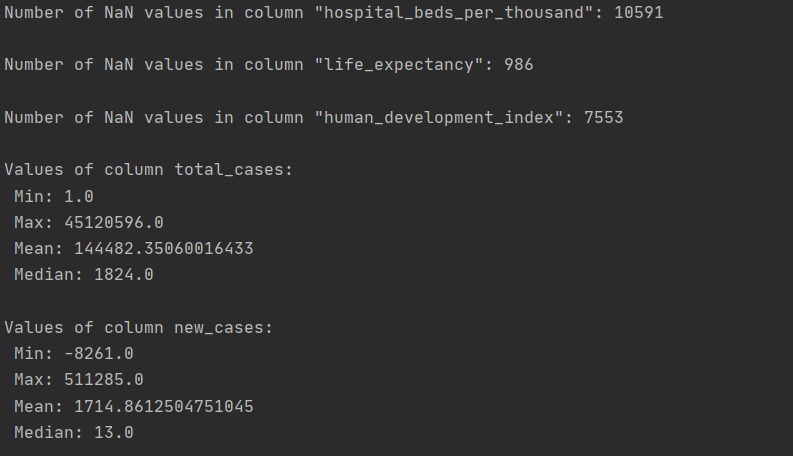
شکل 1 - نصب کتابخانه‌های مورد نیاز

با استفاده از دستور موجود در شکل 2، کد اجرا شده و نتایج تحلیل، نمایش داده خواهد شد. همچنین در صورت نیاز به دریافت آخرین تسخه مجموعه‌داده، فایل csv موجود را پاک و کد را اجرا کرده تا جدیدترین نسخه مجموعه داده دریافت شود.



شکل 2- اجرای کد

لازم به توضیح است که پس از اجرای کد، نمودار‌ها در پوشه figures و موارد مربوط به گزارش در خروجی چاپ شده توسط برنامه (شکل 3) موجود است. همچنین در صورت استفاده از ابزار [PyCharm](https://downloadly.ir/software/programming/pycharm-1/)[[2]](#footnote-2) نسخه حرفه‌ای، نمودار‌ها در بخش SciVIew نیز به نمایش در خواهد آمد و در صورت اجرا از طریق ترمینال، برای هر نمودار پنجره‌ای مجزا جهت مشاهده نمودار‌ها به نمایش در خواهد آمد.



شکل 3 - بخشی‌ای از خروجی کد

# دید‌‌کلی داده‌ها

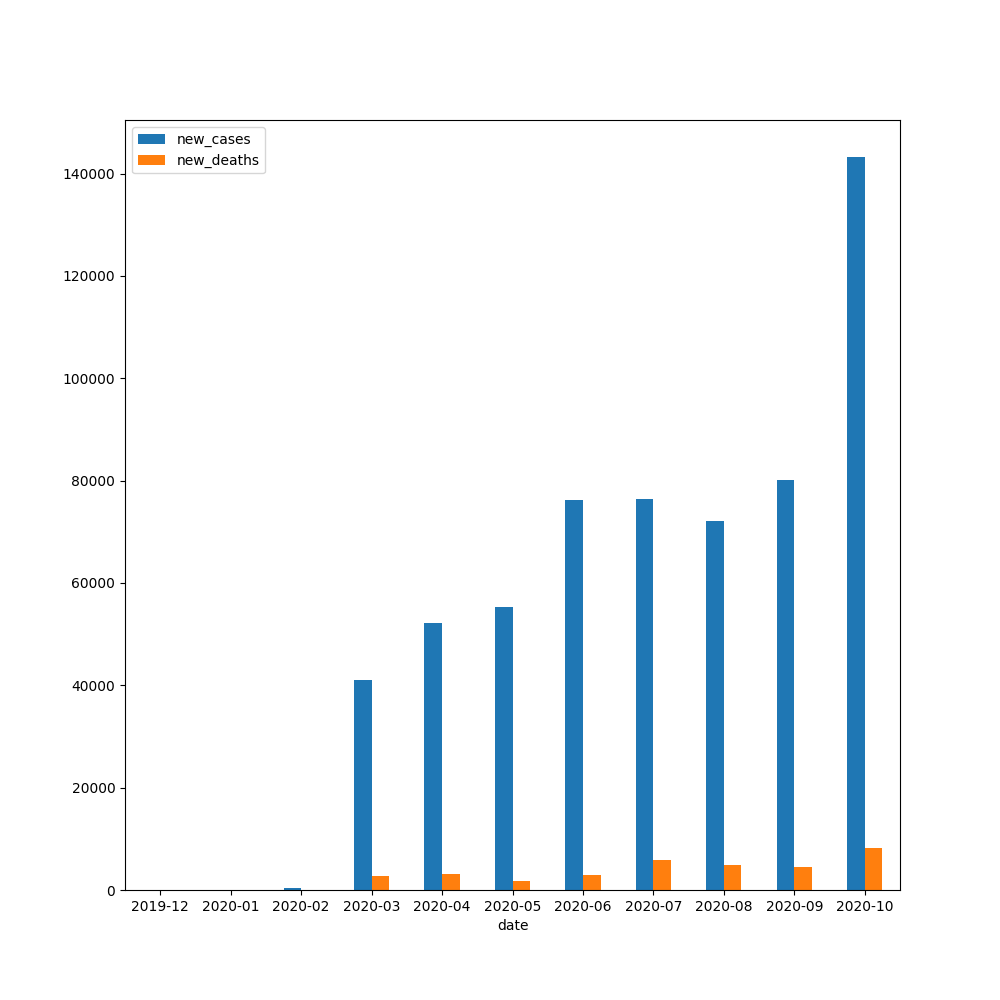
با اجرای کد، نتایج موجود در جدول 1 بدست آمد.

جدول 1- بررسی 10 فیلد از مجموعه‌داده

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| نام فیلد | تعداد سطر‌های بدون مقدار | مقدار بیشینه | نام کشور با مقدار بیشینه | مقدار کمینه | نام کشور با مقدار کمینه | میانه | میانگین | تعداد مقادیر غیر‌معتبر (منفی) |
| تمامی موارد (Total Cases) | 3627 | 45120596 | کل دنیا | 1 | افغانستان | 1824.0 | 14482.35 | 0 |
| موارد جدید (New Cases) | 910 | 511285 | کل دنیا | 8261- | اکوادور | 13 | 1714.86 | 18 |
| کل کشته‌شدگان (Total Deaths) | 12628 | 1182408 | کل دنیا | 1 | آروبا | 76 | 6334.35 | 0 |
| تعداد کشته‌شدگان جدید (New Deaths) | 910 | 10491 | کل دنیا | 1918- | اسپانیا | 0 | 44.93 | 8 |
| تمامی موارد بر میلیون | 3896 | 59108.264 | آندورا | 0.001 | هند | 515.51 | 2664.42 | 0 |
| موارد جدید بر میلیون | 974 | 8652.658 | واتیکان | 2212.545- | لوکزامبورگ | 1.97 | 32.61 | 17 |
| تعداد کشته‌شدگان بر میلیون | 12882 | 1237.55 | سان مارینو | 0 | کل دنیا | 18.55 | 87.92 | 0 |
| تعداد کشته‌شدگان جدید بر میلیون | 974 | 63.14 | سان مارینو | 9.67- | قرقیزستان | 0.026 | 0.6 | 8 |
| تعداد کل تست‌ها (Total Test) | 33178 | 160000000 | چین | 1 | دانمارک | 241845.5 | 2027118.9 | 0 |
| تست‌های جدید | 33336 | 1492409 | هند | 3743- | اکوادور | 3519 | 26509.55 | 7 |

# بررسی موارد جدید و کشته‌شدگان ایران در بازه‌های زمانی مختلف

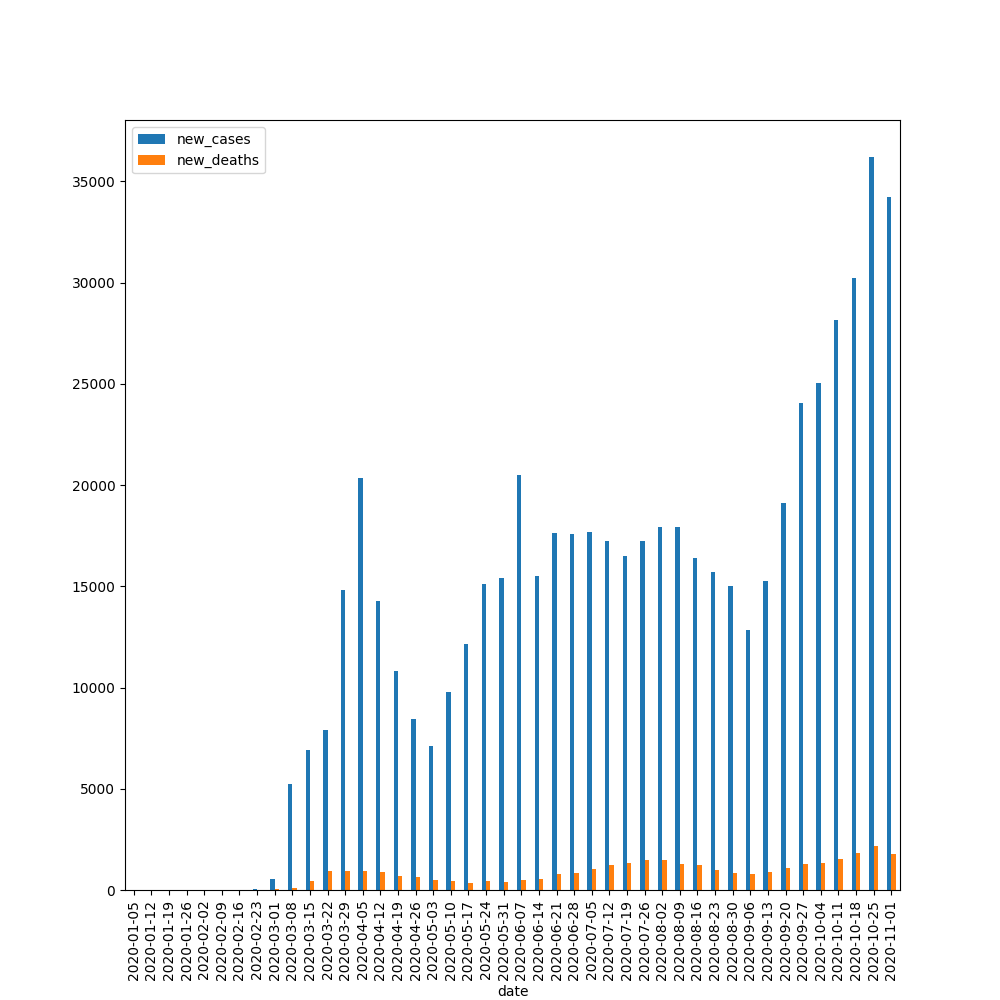
در شکل 4 موارد جدید و کشته‌شده به صورت ماهانه نمایش داده شده است.



شکل 4 - نمودار میله‌ای موارد و کشته‌شدگان جدید در ایران در به صورت ماهانه

در شکل 5، موارد جدید و کشته‌شده در به صورت هفتگی نمایش داده شده است. لازم به ذکر است که در محور افقی، تاریخ‌های ذکر شده، تاریخ شروع هر هفته است و مقادیر هر تاریخ جمع مقادیر آن تاریخ و 6 روز بعد از آن است. همچنین شروع هر هفته یک‌شنبه در نظر گرفته شده است.

به علت بزرگ بودن نمودار موارد به صورت روزانه امکان گنجاندن آن در این گزارش فراهم نبوده و همراه این گزارش فایل این نمودار به صورت مجزا با نام Q2\_daily\_cases.png قابل دسترس است.



شکل 5 - نمودار میله‌ای موارد و کشته‌شدگان جدید در ایران در به هفتگی

# بررسی داده‌های گم‌شده

به صورت کلی دو روش برای مواجه با داده‌ی گم‌شده را می‌توان نام برد:

* **پاک‌کردن داده**: تمام داده‌های ناقص یا غلط پاک شود، این حالت برای زمانی که تعداد داده‌های موجود به اندازه کافی موجود است، می‌توان گفت که مناسب بوده و می‌توان بدون از دست‌دادن نکته خاصی، اینکار را انجام داد اما موضوع مورد توجه این است که باید داده‌ی گم شده، به صورت تصادفی و بدون هیچ‌گونه الگوی خاصی گم شده باشد.
* **بازیابی مقدار**: با استفاده از برخی روش‌های آماری (مثل میانگین، استفاده از متداول‌ترین مقدار، استفاده از رگرسیون[[3]](#footnote-3) و غیره ) می‌توان برخی مقادیر گم‌شده را حدس زد.

یک روش پیشنهادی می‌تواند به این صورت باشد که در مواجه با ستونی که داده‌ی غلط یا گم‌شده دارد از میانگین استفاده شود، برای دقیق‌تر کردن آن می‌توان از میانگین موارد جدید در هفته‌ی داده گم‌شده استفاده کرد.

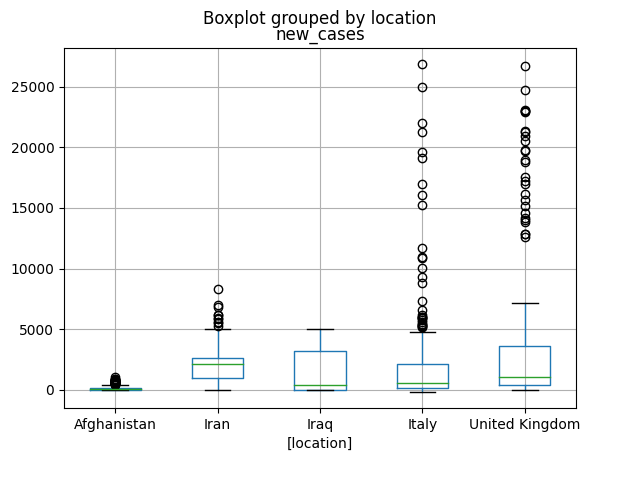
یک حالت دیگر می‌تواند این باشد که فرض شود بخشی از این اعداد منفی توسط کاربر، به صورت اشتباه منفی وارد شده است و می‌توان مقدار مثبت آن ستون را جایگزین مقدار غلط کرده و یک بررسی‌ای انجام شود که آیا با وجود عدد مثبت، تقریبا از الگوی افزایش یا کاهش موجود در هفته یا ماه یا میان دو روز پیروی می‌کند یا خیر.

با توجه به اینکه در داده‌های کشور فرانسه فقط یک سطر دارای مقدار غلط می‌باشد، روش‌های ساده‌ای مثل حذف داده نیز جوابگو خواهد بود اما برای تطبیق پذیری بهتر پیشنهاد می‌شود از میانگین دو روز قبل و بعد این تاریخ استفاده شود.

با توجه به خروجی کد و توضیحات ارائه شده مقدار پیشنهادی برای سطر با مقدار غلط 345 می‌باشد.

# بررسی نمودار‌ جعبه‌ای ایران و چهار کشور انتخابی

با توجه به شکل 6، برای ایران و چهار کشور عراق، افغانستان، انگلستان و ایتالیا نمودار جعبه‌ای[[4]](#footnote-4) تهیه و موارد جدید در آن گنجانده شد.



شکل 6 - نمودار جعبه‌ای کشور‌های ایران، افغانستان، عراق، انگلستان، ایتالیا

## بررسی مقادیر محاسبه‌شده برای کشور ایران

با توجه به خروجی کد، موارد موجود در جدول 2 برای کشور ایران محاسبه گردید.

جدول 2- مقادیر محاسبه‌شده برای کشور ایران

|  |  |
| --- | --- |
| مقدار متغیر | نام متغیر |
| 976 | Q1 |
| 2619 | Q3 |
| 1643 | IQR |
| 5039 | Top Whisker |
| 0 | Bottom Whisker |

## بررسی داده‌های پرت

با توجه به شکل 6 و داده‌های جدول 2، 10 داده‌های پرت (بزرگتر از Top whisker) موجود در ، بدست آمد.

جدول 3 - 10 مورد از داده‌های پرت

|  |  |
| --- | --- |
| موارد جدید | تاریخ |
| 8293 | 2020-10-30 |
| 6968 | 2020-10-28 |
| 6824 | 2020-10-29 |
| 6191 | 2020-10-26 |
| 6134 | 2020-10-24 |
| 5960 | 2020-10-27 |
| 5814 | 2020-10-25 |
| 5616 | 2020-10-22 |
| 5471 | 2020-10-23 |
| 5275 | 2020-04-05 |

# بررسی نمودار جعبه‌ای کشور ایران برای موارد هفتگی جدید ابتلا

به علت حجم بالای نمودار، فایل این نمودار با نام bonus\_question\_1\_iran\_new\_cases\_box\_whisker.jpg همراه با این گزارش ارائه می‌شود.

1. Virtual Enviroment [↑](#footnote-ref-1)
2. https://downloadly.ir/software/programming/pycharm-1 [↑](#footnote-ref-2)
3. Regression [↑](#footnote-ref-3)
4. Box Whisker Plot [↑](#footnote-ref-4)