**مستندات پروژه‌ی نهایی**

****

**بوت‌کمپ تحلیل داده کوئرا**

**زمستان 1401**

**آخرین ویرایش: ۲۲ بهمن 1401**

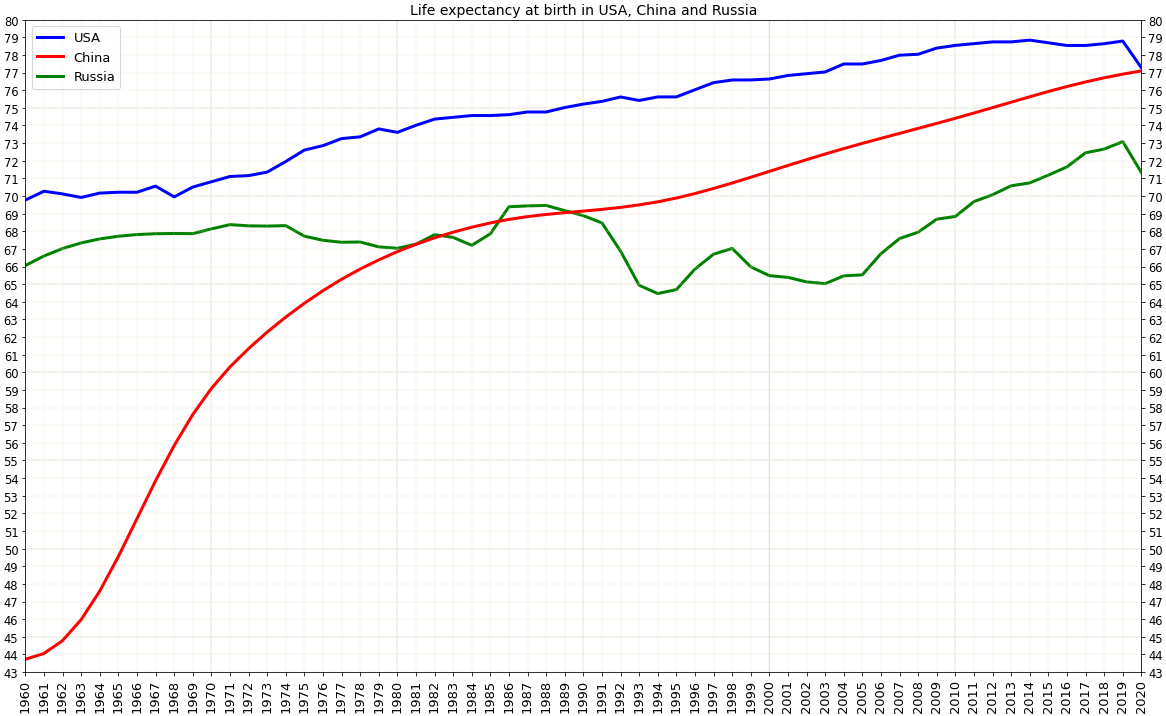
**مقدمه**

[امید به زندگی](https://fa.wikipedia.org/wiki/%D8%A7%D9%85%DB%8C%D8%AF_%D8%A8%D9%87_%D8%B2%D9%86%D8%AF%DA%AF%DB%8C) یا انتظار زندگی[[1]](#footnote-1) یک معیار آماری است که متوسط زمانی را که انتظار می‌رود یک موجود زنده زندگی کند نشان می‌دهد. این معیار می‌تواند از عوامل مختلفی نظیر سال تولد، سن کنونی، جنسیت، کشور، وضعیت اقتصادی و بهداشتی و غیره تاثیر پذیرد. انتظار زندگی را می‌توان در حالت‌های مختلفی سنجید که رایج‌ترین آن‌ها عبارتند از:

|  |  |
| --- | --- |
| * انتظار زندگی در لحظه‌ی تولد | * انتظار زندگی در 15 سالگی |
| * انتظار زندگی در 65 سالگی | * انتظار زندگی در 80 سالگی |

در این پروژه متغیر هدف ما *انتظار زندگی در لحظه‌ی تولد* است. یعنی هنگامی‌که انسانی به دنیا می‌آید انتظار می‌رود به‌طور میانگین چند سال عمر کند.

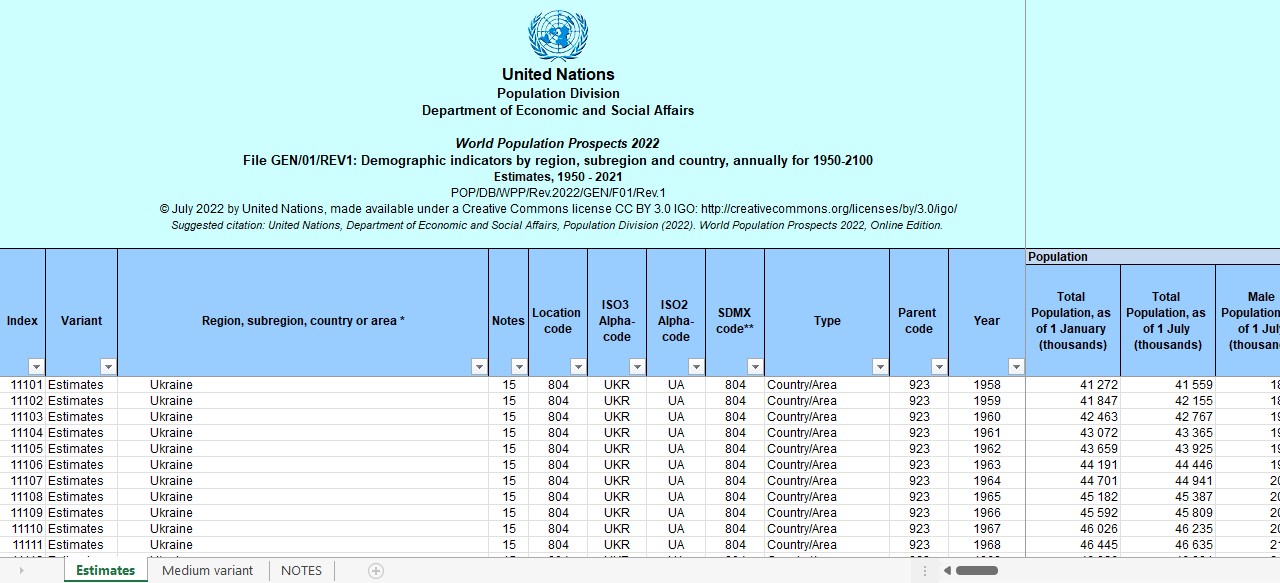
برای درک شهودی‌تر، نموداری از انتظار زندگی در لحظه‌ی تولد برای سه کشور ایالات متحده‌ی آمریکا، چین و روسیه در سال‌های 1960 تا 2020 در شکل زیر آورده شده است. همان‌طور که مشاهده می‌کنید انتظار زندگی در کشور چین در طول این سال‌ها همواره صعودی بوده و توانسته در سال 2020 بسیار به ایالت متحده‌ی آمریکا نزدیک شود!



در این پروژه از یک مجموعه‌داده‌ی استاندارد استفاده کرده و تحلیل‌های جذابی بر روی داده‌های آن انجام خواهیم داد. البته قرار است خودتان نیز به سراغ جمع‌آوری داده‌‎های کاربردی بیشتر بروید! علاوه‌براین در انتها از شما می‌خواهیم به کمک یادگیری ماشین مدلی ارائه کنید که با توجه به عوامل مختلف و گزارش‌ سال‌های پیشین قادر باشد تخمینی از امید به زندگی در سال‌های آینده را ارائه دهد.

**مجموعه‌داده**

در این پروژه از داده‌های گزارش جمعیت جهان[[2]](#footnote-2) که توسط سازمان ملل متحد[[3]](#footnote-3) تهیه شده استفاده خواهیم کرد. این مجموعه‌داده شامل گزارش این سازمان برای نواحی مختلف جهان در سال‌های 1950 تا 2021 و تخمین آن برای سال‌های بعدی است. منظور از نواحی، گزارش به شکل جهانی، قاره‌ای، کشوری و غیره است که به کمک ستون Type می‌توان نوع آن را تشخیص داد.



این مجموعه‌داده به شکل یک فایل اکسل در اختیار شما قرار گرفته که شامل برگه‌های[[4]](#footnote-4) مختلفی است. در برگه‌ی Estimates اطلاعات سال‌های پیشین (تا 2021) و در برگه‌ی Medium variant پیش‌بینی سال‌های بعدی قرار گرفته است.

هر ردیف با ویژگی‌های مختلفی شامل آمارهای جمعیتی[[5]](#footnote-5)،‌ باروری[[6]](#footnote-6)، مرگ و میر [[7]](#footnote-7)و مهاجرت[[8]](#footnote-8) همراه است. انتظار زندگی در لحظه‌ی تولد نیز در ستونی به نام Life Expectancy at Birth, both sexes (years) قرار گرفته است.

با این حال، عوامل تاثیرگذار بسیار مهم دیگری بر روی انتظار زندگی وجود دارند که در این مجموعه‌داده آورده نشده‌اند. از شما انتظار می‌رود با توجه به ایده‌ها و تحلیل‌های خودتان نسبت به جمع‌آوری داده‌های مربوط به این عوامل و ادغام آن‌ها با مجموعه‌داده‌ی اصلی اقدام کنید. لیستی از عوامل پیشنهادی در زیر آورده شده است (با این حال، ایده‌پردازی خود را محدود به این لیست نکنید):

* تولید ناخالص ملی[[9]](#footnote-9) (این مورد برای تمام تیم‌ها ضروری است)
  + می‌توانید ترجیحاً از GDP per Capita استفاده کنید ([مطالعه‌ی تفاوت این دو معیار](https://pediaa.com/what-is-the-difference-between-gdp-and-gdp-per-capita/))
* نسبت هزینه‌های درمانی-بهداشتی دولت به کل هزینه‌های دولت
* میانگین شاخص توده‌ی بدنی[[10]](#footnote-10) کل جمعیت کشور
* میزان مصرف الکل
* میزان واکسناسیون (مثلاً درصد افراد ایمن‌شده در برابر هپاتیت ب،‌ دیفتری، فلج اطفال و غیره)
* تعداد موارد مبتلا به سرخک
* تعداد مبتلایان به HIV/AIDS یا فوتی‌های آن
* وضعیت توسعه‌یافتگی کشور
* وضعیت تحصیلات مردم کشور

معرفی چند منبع کمکی:

|  |  |  |  |
| --- | --- | --- | --- |
| * [منبع 1](https://clio-infra.eu/Indicators/LifeExpectancyatBirthTotal.html) | * [منبع 2](https://databank.worldbank.org/databases) | * [منبع 3](https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data) |  |

نکته: توجه داشته باشید که درست است این قسمت به‌عنوان نخستین گام پروژه آورده شده اما ممکن است در طول پروژه (خصوصاً در بخش یادگیری ماشین) تصمیم بگیرید ویژگی‌های بیشتری را به مجموعه‌ی خود اضافه کنید.

نکته: تصمیم‌گیری درباره‌ی نحوه‌ی ذخیره‌سازی داده‌ها بر عهده‌ی خودتان است. در صورت علاقه می‌توانید برای آن یک دیتابیس تعریف کنید یا مستقیماً به کمک pandas بر روی فایل اکسل یا csv کار کنید.

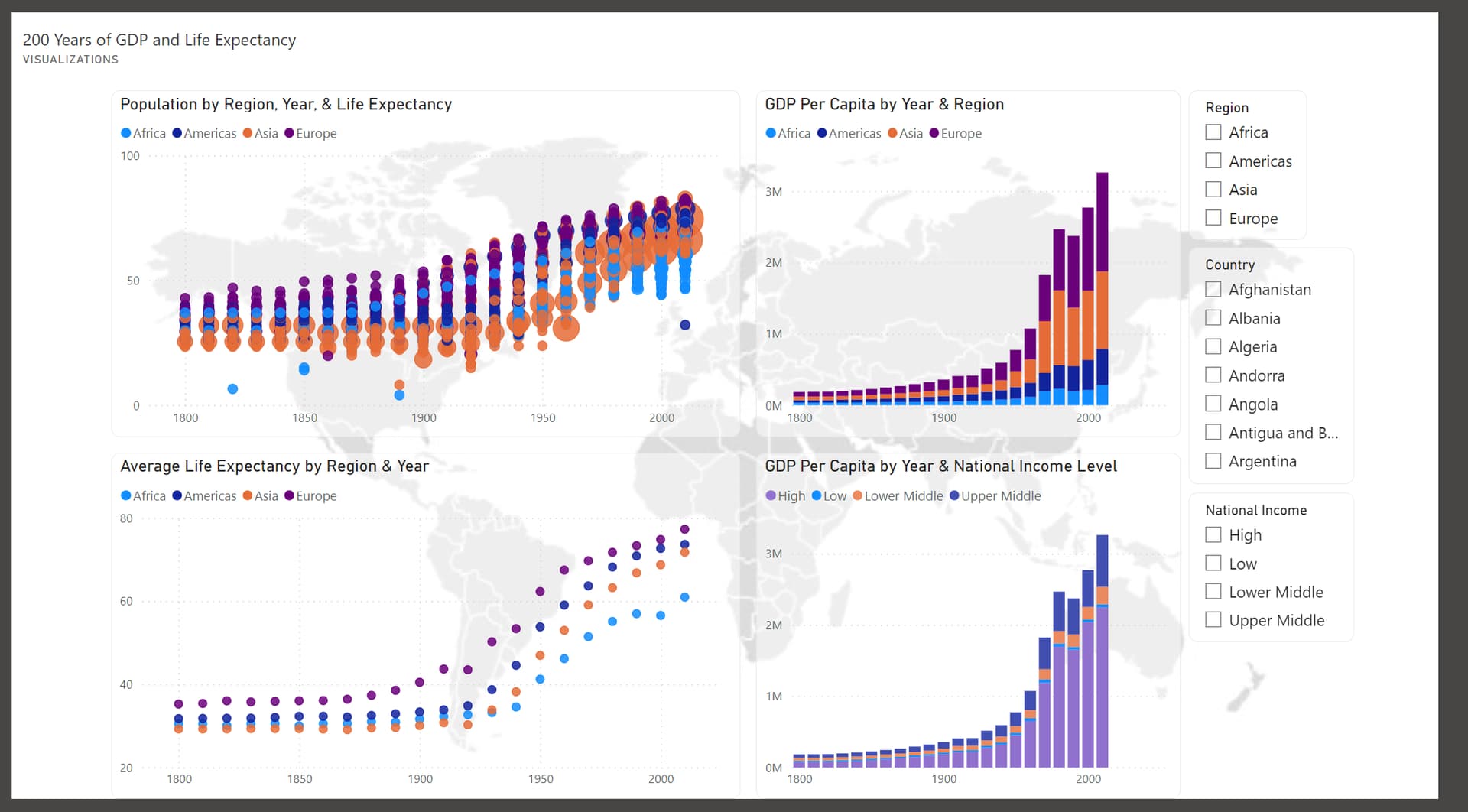
توجه: نیاز است به غیر از تولید ناخالص ملی (GDP) حداقل 5 ویژگی دیگر را به مجموعه‌داده اضافه کنید.

**تحلیل داده**

در این قسمت قصد داریم به چندین پرسش مهم بر روی داده‌های خود پاسخ داده و به کمک ابزارهای داشبورد‌ساز (همچون Power BI) تحلیل‌ها و تفسیرهای مفیدی را از داده‌ها ارائه دهیم.

در داشبوردی که طراحی می‌کنید نیاز است به تمام پرسش‌های زیر پاسخ داده شود:

1. نموداری از میانگین انتظار زندگی برای هر قاره براساس سال همچون تصویر زیر:



1. نموداری از انتظار زندگی هر کشور براساس هر سال همچون تصویر ابتدای پروژه (با امکان انتخاب نمایش تنها کشورهای مربوط به یک قاره)
2. نمایش 10 کشوری که از سال 2000 تا 2021 بیشترین افزایش انتظار زندگی را داشته‌اند
   * یعنی اگر در سال 2000 انتظار زندگی کشور و در سال 2021 انتظار زندگی آن باشد، باید 10 کشوری که بیشترین را داشته‌اند گزارش دهید.
3. نمودار انتظار زندگی در لحظه‌ی تولد براساس سال و سطح درآمد
   * برای سطح درآمد به ایندکس شماره‌ی 1155 با عنوان World Bank income groups دقت کنید.
4. نمودار تولد توسط زنان بین 15 تا 19 سال براساس سطح درآمد و سال. همچنین گزارش 10 کشوری که در مجموع سال‌های 2000 تا 2021 بیشترین فرزندآوری در این سنین را داشته‌اند.
5. نمودار GDP per Capita هر قاره براساس هر سال
6. نمایش میانگین انتظار زندگی هر کشور در سال‌های 2000 تا 2021 بر روی نقشه‌ی زمین
   * می‌توانید به کمک رنگ، اندازه یا شدت اطلاعات مفیدی را روی نقشه نمایش دهید
7. گزارشی از تفاوت انتظار زندگی براساس جنسیت (نوع نمودار یا گزارش بر عهده‌ی خودتان)
8. گزارشی جهانی براساس هر سال شامل:
   * جمعیت کلی جهان
   * تعداد مردان و زنان
   * تعداد افراد متولد شده
   * تعداد افراد فوت کرده
   * میانگین انتظار زندگی در لحظه‌ی تولد
   * میزان افزایش یا کاهش انتظار زندگی در لحظه‌ی تولد نسبت به سال گذشته
   * (و موارد بیشتر در صورت علاقه‌ی خودتان)

نکته: علاوه بر این موارد می‌توانید با توجه به ایده‌های خودتان یا ویژگی‌های بیشتری که استخراج کرده‌اید تحلیل‌های جذاب و مفید دیگری را به‌عنوان نمره‌ی اضافه ارائه دهید. همچنین می‌توانید از تحلیل‌های موجود در [این صفحه](https://ourworldindata.org/life-expectancy) نیز ایده بگیرید.

**مدل‌سازی (یادگیری ماشین)**

اکنون نیاز است به کمک یادگیری ماشین، مدلی آموزش دهید که بتواند با دریافت مقادیر ویژگی‌های مختلف نسبت به پیش‌بینی انتظار زندگی در لحظه‌ی تولد برای آن اقدام کند. نمونه‌های آموزشی برای سال‌های پیشین در مجموعه‌داده‌ی اصلی (برگه‌ی Estimates) در اختیارتان قرار گرفته و ویژگی‌های مفید بیشتری را نیز خودتان به مجموعه اضافه کرده‌اید. حال باید مدل شما بتواند با دریافت مقادیر این ویژگی‌ها برای سال 2022 پیش‌بینی‌هایی تولید کند که بسیار نزدیک به تخمین سازمان ملل که در برگه‌ی Medium variant آمده باشد.

نوع مدل انتخابی، مراحل پیش‌پردازش ویژگی‌ها، نوع ارزیابی و تمام موارد دیگر کاملاً برعهده‌ی خودتان و براساس یادگیری شما از این مبحث است.

نیاز است مدل نهایی شما مقادیر قابل‌قبولی برای سال 2022 تولید کند. برای مقایسه‌ی عملکرد مدل خود با تخمین ارائه‌شده توسط سازمان ملل یک معیار پیشنهادی می‌تواند باشد.

موفق باشید

1. [Life expectancy](https://en.wikipedia.org/wiki/Life_expectancy) [↑](#footnote-ref-1)
2. World Population Prospects [↑](#footnote-ref-2)
3. United Nations [↑](#footnote-ref-3)
4. sheets [↑](#footnote-ref-4)
5. population [↑](#footnote-ref-5)
6. fertility [↑](#footnote-ref-6)
7. mortality [↑](#footnote-ref-7)
8. migration [↑](#footnote-ref-8)
9. Gross Domestic Product (GDP) [↑](#footnote-ref-9)
10. Body Mass Index (BMI) [↑](#footnote-ref-10)