

مستندات پروژه‌ی نهایی



بوت‌کمپ تحلیل داده کوئرا

زمستان ۱۴۰۱

آخرین ویرایش: ۲۲ بهمن ۱۴۰۱

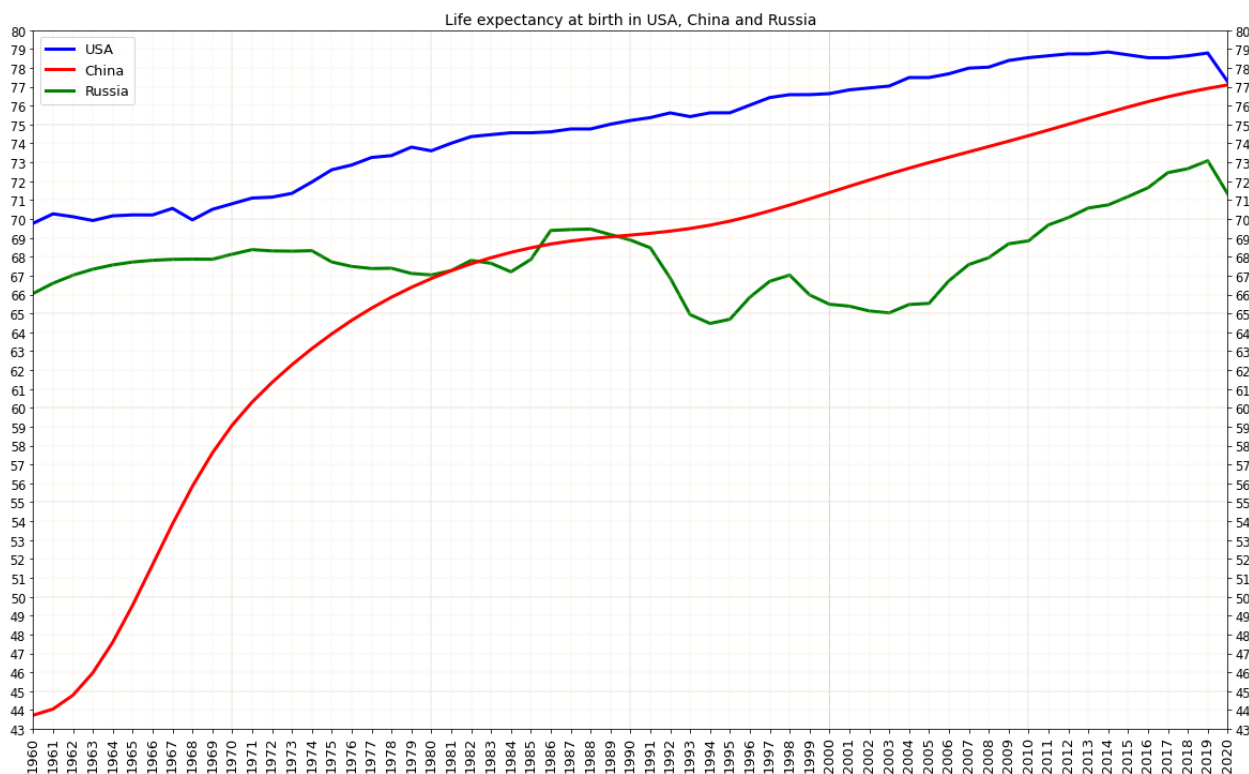
مقدمه

امید به زندگی یا انتظار زندگی^۱ 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 ۶۵ سالگی
- انتظار زندگی در ۱۵ سالگی
- انتظار زندگی در ۸۰ سالگی

در این پروژه متغیر هدف ما *انتظار زندگی در لحظه‌ی تولد* است. یعنی هنگامی که انسانی به دنیا می‌آید انتظار می‌رود به طور 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 ۱۹۶۰ تا ۲۰۲۰ 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 ۲۰۲۰ بسیار به ایالت متحده‌ی آمریکا نزدیک شود!




¹ [Life expectancy](#)

در این پروژه از یک مجموعه داده‌ی 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.

مجموعه داده

در این پروژه از داده‌های گزارش جمعیت جهان^۲ که توسط سازمان ملل متحد^۳ تهیه شده استفاده خواهیم کرد. این مجموعه داده شامل گزارش این سازمان برای نواحی مختلف جهان در سال‌های ۱۹۵۰ تا ۲۰۲۱ 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.



United Nations
 Population Division
 Department of Economic and Social Affairs

World Population Prospects 2022
 File GEN/01/REV1: Demographic indicators by region, subregion and country, annually for 1950-2100
 Estimates, 1950 - 2021
 POP/DB/WFP/Rev2022/GEN/F01/Rev.1
 © July 2022 by United Nations, made available under a Creative Commons license CC BY 3.0 IGO: <http://creativecommons.org/licenses/by/3.0/igo/>
 Suggested citation: United Nations, Department of Economic and Social Affairs, Population Division (2022). World Population Prospects 2022, Online Edition.

Index	Variant	Region, subregion, country or area *	Notes	Location code	ISO3 Alpha-code	ISO2 Alpha-code	SDMX code**	Type	Parent code	Year	Population		
											Total Population, as of 1 January (thousands)	Total Population, as of 1 July (thousands)	Male Population of 1 Jul (thousan
11101	Estimates	Ukraine	15	804	UKR	UA	804	Country/Area	923	1958	41 272	41 559	18
11102	Estimates	Ukraine	15	804	UKR	UA	804	Country/Area	923	1959	41 847	42 155	18
11103	Estimates	Ukraine	15	804	UKR	UA	804	Country/Area	923	1960	42 463	42 767	18
11104	Estimates	Ukraine	15	804	UKR	UA	804	Country/Area	923	1961	43 072	43 365	18
11105	Estimates	Ukraine	15	804	UKR	UA	804	Country/Area	923	1962	43 659	43 925	18
11106	Estimates	Ukraine	15	804	UKR	UA	804	Country/Area	923	1963	44 191	44 446	18
11107	Estimates	Ukraine	15	804	UKR	UA	804	Country/Area	923	1964	44 701	44 941	20
11108	Estimates	Ukraine	15	804	UKR	UA	804	Country/Area	923	1965	45 182	45 387	20
11109	Estimates	Ukraine	15	804	UKR	UA	804	Country/Area	923	1966	45 592	45 809	20
11110	Estimates	Ukraine	15	804	UKR	UA	804	Country/Area	923	1967	46 026	46 235	20
11111	Estimates	Ukraine	15	804	UKR	UA	804	Country/Area	923	1968	46 445	46 635	20

این مجموعه داده به شکل یک فایل اکسل در اختیار شما قرار گرفته که شامل برگه‌های^۴ مختلفی است. در برگه‌ی Estimates اطلاعات سال‌های پیشین (تا ۲۰۲۱) و در برگه‌ی Medium variant پیش‌بینی سال‌های بعدی قرار گرفته است.

² World Population Prospects

³ United Nations

هر ردیف با ویژگی‌های مختلفی شامل آمارهای جمعیتی^۵، باروری^۶، مرگ و میر^۷ و مهاجرت^۸ 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 مجموعه داده 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 مجموعه داده‌ی 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^۹ (این مورد برای تمام تیم‌ها **ضروری** است)
 - می‌توانید ترجیحاً از GDP per Capita استفاده کنید ([مطالعه‌ی تفاوت این دو معیار](#))
 - نسبت هزینه‌های درمانی-بهداشتی دولت به کل هزینه‌های دولت
 - میانگین شاخص توده‌ی بدنی^{۱۰} کل جمعیت کشور
 - میزان مصرف الکل
 - میزان واکسناسیون (مثلاً درصد افراد ایمن‌شده در برابر هپاتیت ب، دیفتی، 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 ۳](#)

❖ [منبع ۲](#)

❖ [منبع ۱](#)

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.

⁵ population

⁶ fertility

⁷ mortality

⁸ migration

⁹ Gross Domestic Product (GDP)

¹⁰ Body Mass Index (BMI)

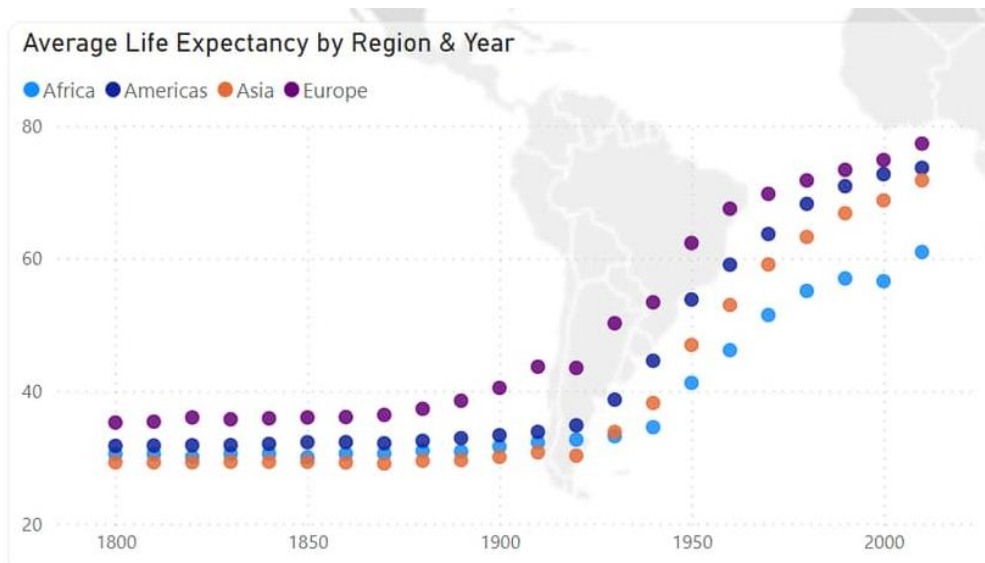
توجه: نیاز است به غیر از تولید ناخالص ملی (GDP) حداقل ۵ ویژگی دیگر را به مجموعه داده اضافه کنید.

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:

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



۲. نموداری از انتظار زندگی هر کشور براساس هر سال همچون تصویر ابتدای پروژه (با امکان انتخاب نمایش تنها کشورهای مربوط به یک قاره)

۳. نمایش ۱۰ کشوری که از سال ۲۰۰۰ تا ۲۰۲۱ بیشترین افزایش انتظار زندگی را داشته‌اند

- یعنی اگر در سال ۲۰۰۰ انتظار زندگی کشور x و در سال ۲۰۲۱ انتظار زندگی آن y باشد، باید ۱۰ کشوری که بیشترین $y - x$ را داشته‌اند گزارش دهید.

۴. نمودار انتظار زندگی در لحظه‌ی تولد براساس سال و سطح درآمد

- برای سطح درآمد به ایندکس شماره‌ی ۱۱۵۵ با عنوان World Bank income groups دقت کنید.

۵. نمودار تولد توسط زنان بین ۱۵ تا ۱۹ سال براساس سطح درآمد و سال. همچنین گزارش ۱۰ کشوری که در مجموع سال‌های ۲۰۰۰ تا ۲۰۲۱ بیشترین فرزندآوری در این سنین را داشته‌اند.

۶. نمودار GDP per Capita هر قاره براساس هر سال

۷. نمایش میانگین انتظار زندگی هر کشور در سال‌های ۲۰۰۰ تا ۲۰۲۱ بر روی نقشه‌ی زمین

- می‌توانید به کمک رنگ، اندازه یا شدت اطلاعات مفیدی را روی نقشه نمایش دهید
- ۸. گزارشی از تفاوت انتظار زندگی براساس جنسیت (نوع نمودار یا گزارش بر عهده‌ی خودتان)
- ۹. 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](#) 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 مجموعه داده‌ی 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 ۲۰۲۲ 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 ۲۰۲۲ تولید کند. برای مقایسه‌ی عملکرد مدل خود با تخمین ارائه‌شده توسط سازمان ملل یک معیار پیشنهادی می‌تواند R^2 Score باشد.

موفق باشید