UDACITY

Return to Classroom

DISCUSS ON STUDENT HUB

Investigate a Dataset

REVIEW

HISTORY

Requires Changes

3 specifications require changes

Hi Student,

Well done in your first submission!. I think the work done is very good: the explanations, as well as the questions and arguments, are clear and well written.

I would like to highlight the way of documenting the project, in a very orderly manner and with enough comments that explain the project process and the observations of the data wrangling and the data cleaning phase. It is also worth highlighting the large number of visualizations that have been included in the analysis. It shows that you have put a lot of effort into doing this project. Great job!.

As you can see, some sections need clarification to meet the rubric specifications. Note that this is a first review, so it is normal to have to refine certain aspects of the project. :)

I summarize the points to fix:

- To avoid repetitive code, please, try to create user defined functions.
- Remember to comment the limitations encountered in the dataset in the Conclusions section.
- Try to include an explanation about each set of plots.

I hope the explanations and resources I have given in this review are clear enough to fit the requirements.
Please, if you still have any questions don't hesitate to ask in **Knowledge** where mentors will be glad to help you.

Good luck with your next submission!.

Nice job so far!! :)

Code Functionality

All code is functional and produces no errors when run. The code given is sufficient to reproduce the results described.

Awesome:

- You have done very well by including the file created from a Jupyter notebook, this is crucial for reviewers to check that all the code cells execute correctly. In your case, the code is good and with no errors when it runs. I like the clean coding style and documentation. Good job!.

Learning Note:

- If you want to delve into the Jupyter environment, here you have a very interesting post about **How to get more productivity in the notebook environment**. I think it could be very interesting to you.

The project uses NumPy arrays and Pandas Series and DataFrames where appropriate rather than Python lists and dictionaries. Where possible, vectorized operations and built-in functions are used instead of loops.

Awesome:

- The numpy, pandas and Dataframes used are really good. I like the use of the different tools like `value_counts`, `sort_values` and the simplicity of some solutions. In addition, you have also used `.info()` and `.describe()` functions to study the structure of the dataset. Using Pandas helps to shorten the procedure of handling data and provides a huge set of important commands and features which are used to easily analyze the data. **Well done!.**

Learning Note:

- Regarding Pandas, I think that probably this is **one of the best python pandas tutorials available**, it is specifically useful for people working with data cleansing and analysis, and **this one** and **this one** are two fantastic posts about how to prepare and analyze a dataset with pandas.

The code makes use of functions to avoid repetitive code. The code contains good comments and variable names, making it easy to read.

Awesome:

- Well done by naming each variable appropriately and by including the comments that describe each

part of the code when necessary. This way for any programmer who reads your code will be easy to understand.

Required: ❗

- Please, note that in this section it is needed (by rubric) the use of functions to avoid repetitive code: *"The code makes use of functions to avoid repetitive code."*.

  The Don't Repeat Yourself (DRY) principle in software engineering is a fantastic mantra that we need to keep in mind in each of our projects (big or little, it doesn't matter). This principle is studied and discussed in a lot of posts, but I'd like to share with you **this one** which I think is especially well explained.

  For example, in your project, **you could have created a function in the cells In[62], In[65], In[73], In[79], In[87] and In[90] that draws the bar plots passing as argument the data, title and labels**, Why?, because if you decide to change the way you are plotting your graphs, you only will need to change that function instead of in each place where you have plotted a graph. This is the main idea for the DRY principle!.

  Some of the code cells in your project I have referred to before:

In [62]:
```
#plotting top 5 accross the years
united_states = top_prod.iloc[0, :-2]
united_states.plot(kind='bar', title='Energy Production in United States')
plt.xlabel('Years')
plt.ylabel('Energy Produced (Tonnes of Oil Equivalent)')
plt.show()

china = top_prod.iloc[1, :-2]
china.plot(kind='bar', title='Energy Production in China')
plt.xlabel('Years')
plt.ylabel('Energy Produced (Tonnes of Oil Equivalent)')
plt.show()

russia = top_prod.iloc[2, :-2]
russia.plot(kind='bar', title='Energy Production in Russia')
plt.xlabel('Years')
plt.ylabel('Energy Produced (Tonnes of Oil Equivalent)')
plt.show()

saudi_arabia = top_prod.iloc[3, :-2]
saudi_arabia.plot(kind='bar', title='Energy Production in Saudi Arabia')
plt.xlabel('Years')
plt.ylabel('Energy Produced (Tonnes of Oil Equivalent)')
plt.show()

india = top_prod.iloc[4, :-2]
india.plot(kind='bar', title='Energy Production in India')
plt.xlabel('Years')
plt.ylabel('Energy Produced (Tonnes of Oil Equivalent)')
plt.show()
```

In [65]:
```
singapore = low_prod.iloc[0, :-3]
singapore.plot(kind='bar', title='Energy Production in Singapore')
plt.xlabel('Years')
plt.ylabel('Energy Produced (Tonnes of Oil Equivalent)')
plt.show()

cyprus = low_prod.iloc[1, :-3]
cyprus.plot(kind='bar', title='Energy Production in Cyprus')
plt.xlabel('Years')
plt.ylabel('Energy Produced (Tonnes of Oil Equivalent)')
plt.show()

hong_kong_china = low_prod.iloc[2, :-3]
hong_kong_china.plot(kind='bar', title='Energy Production in Hong Kong, China')
plt.xlabel('Years')
plt.ylabel('Energy Produced (Tonnes of Oil Equivalent)')
plt.show()

luxembourg = low_prod.iloc[3, :-3]
luxembourg.plot(kind='bar', title='Energy Production in Luxembourg')
plt.xlabel('Years')
plt.ylabel('Energy Produced (Tonnes of Oil Equivalent)')
plt.show()

moldova = low_prod.iloc[4, :-3]
moldova.plot(kind='bar', title='Energy Production in Moldova')
plt.xlabel('Years')
plt.ylabel('Energy Produced (Tonnes of Oil Equivalent)')
plt.show()
```

Learning notes:

- I would like to share with you these fantastic posts about **Why to use functions?** and **Why to use**

- I would like to share with you these fantastic posts about Why to use functions? and Why to use comments?.
- Here you have a tutorial about how to create functions in Python, that I think could be very useful to you.

Quality of Analysis

The project clearly states one or more questions, then addresses those questions in the rest of the analysis.

Awesome:

- The questions are well stated and well explained, they have been stated in the introduction section so that the audience has a clear understanding of what this analysis is about. The answers are clear and well written.
  It is very important in a work environment, where you write your analysis for other people, to clarify what questions are going to be addressed and the data used, so the reader can have a clear understanding of what the project is about from the very beginning.**Good job!.**

  Questions stated in your Introduction section:

  Introduction

  In this project, I'll be analysing data associated with total energy produced and consumed in different countries over about two decades, and the Consumption CO2 per capita for the people in the countries.

  Project Aim

  The main aim of this project it to explore trends on energy production, consumption and CO_2 emissions within two decades from around the world. The research questions include:

  - Which countries are the top and least energy producers?
  - Which countries consume the most and least energy?
  - Which countries are the highest and lowest CO_2 emitters?

  > **3 datasets where used:**: * **Energy Production dataset** - `energy_production_total.csv` . Description: Energy production refers to forms of primary energy--petroleum (crude oil, natural gas liquids, and oil from nonconventional sources), natural gas, solid fuels (coal, lignite, and other derived fuels), and combustible renewables and waste--and primary electricity, all converted into tonnes of oil equivalents.
  >
  > ```
  > * Unit of measurement: Tonnes of oil equivalent (toe)
  >
  > * Source: [World Bank, 2010](https://data.worldbank.org/indicator/EG.EGY.PROD.KT.OE)
  > ```
  >
  > - **Energy Consumption dataset** - `'energy_use_per_person.csv` . Description: Energy use refers to use of primary energy before transformation to other end-use fuels, which is equal to indigenous production plus imports and stock changes, minus exports and fuels supplied to ships and aircraft engaged in international transport.
  >   - Unit of measurement: Kg of oil equivalent per capita
  >   - Source: World Bank, 2015
  > - **Consumption CO_2 per capita dataset** - `consumption_emissions_tonnes_per_person.csv` . Description: Per capita carbon dioxide emissions from the fossil fuel consumption, cement production and gas flaring, minus export, plus import during the given year.
  >   - Unit of measurement: Metric tons of CO2 per person
  >   - Source: Gapminder

- I like very much the Introduction section you have included at the beginning of the project where you explain what the project is about and what datasets you are studying. You also have described the main features of each dataset and given the link to each of them. **Excellent work!!.**

Data Wrangling Phase

The project documents any changes that were made to clean the data, such as merging multiple files, handling missing values, etc.

Awesome:

- I like a lot the work done in your data exploration **checking for the shape of the datasets, checking for data types, checking statistics, checking for null values and handling them, dropping some columns, changing formats**, ...etc... This is a crucial step in every Machine Learning project. I'd like to share with you **this fantastic post**, where almost every aspect about the data exploration is touched.

Learning note:

- The most important aspect of Data Wrangling is to clean or transform the data preparing it for analysis. One main issue is having missing data while conducting analysis, which can provide skew/bias results. There are a few methods that Pandas provide to deal with these issues, here you have an interesting post about how to **handling missing values**.

Exploration Phase

The project investigates the stated question(s) from multiple angles. At least three variables are investigated using both single-variable (1d) and multiple-variable (2d) explorations.

Awesome:

- You have done explorations of both unique variable (1d) (by using the describe() method) and multiple variables (2d) (bar plots) to give answers to the questions that you have formulated at the beginning of the project. Good job!.

Learning note:

- **This page** provides an overview of graphing steps and principles and types of graphs that I think could be interesting to read.

The project's visualizations are varied and show multiple comparisons and trends. Relevant statistics are computed throughout the analysis when an inference is made about the data.

At least two kinds of plots should be created as part of the explorations.

Awesome:

- Well done by including **bar charts and pie plots** showing multiple comparisons and trends. You have

also studied relevant statistics in your analysis, this is useful to understand better the dataset and to answer the questions presented, so the exploration phase is very complete. Good job!.

Learning note:

- If you want to delve more into the different types of visualizations, **here you have a guide to selecting the type of the visualization tool**, in this guide you will find a diagram where you can ask " What would I like to show?", and the answer provides a decision tree to follow and determine the kind of visualization of the data you can use.

Conclusions Phase

The results of the analysis are presented such that any limitations are clear. The analysis does not state or imply that one change causes another based solely on a correlation.

Required: ❗

- You have included a conclusions section where you have commented the results of the analysis based on the questions established and It has been commented the main goal of the project. This is a very good section, however, there is a little work to do to meets specifications.
  Please, note that the Conclusions section should have a **'limitations'** subsection: *"The results of the analysis are presented such that any limitations are clear."* where you should write down the limitations of the dataset and what way it may affect the output. Possible examples include null values, if samples doesn't represent the population, if the datasets have missing values, null values, outliers, ... etc.

  Conclusions Section in the project:

  Conclusions

  The main aim of this project was to explore trends on energy production, consumption and CO_2 emissions within two decades (1990-2009) from around the world. To achieve this aim, the following questions were asked:

  - Which countries are the top and least energy producers?
  - Which countries consume the most and least energy?
  - Which countries are the highest and lowest CO_2 emitters?

  Here are the main findings from the 3 datasets which provided information on energy production, consumption and CO_2 emission from around the world:

  > **Two decades of Energy Production**: In this project it was seen that between 1990 and 2009, the top 5 energy producers were the United States, China, Russia, Saudi Arabia, India and Canada. It was also seen that the lowest producers were Singapore, Cyprus, Hong Kong China, Luxembourg and Moldova.

  > **Two decades of Energy Use Per Person**: Interestingly, none of the top 5 energy producers made it to the top 5 energy consumers as the top 5 consumers of Energy were Qatar, Curaçao, Bahrain, Iceland and United Arab Emirates. Canada and United States on the list of Top Energy Consumers were at position 7 and 8 respectively. This suggests that the top 5 energy producers may not be consuming all the energy that they are producing. Also, it is important to restate that this indicator, Energy Use Per person, is a function of the country's population. A suggestion for further studies would be to place consumption figures at par with population figures. Another interesting insight is that all the top energy producers and consumers are either developed countries or high developing countries. Moving now to low energy consumers, the top 5 were Senegal, Haiti, Yemen, Myanmar and Congo Rep - all of which are low developing countries.

  > **Two decades of CO_2 Emissions per Capita Per Person**: The top 5 CO_2 emitters are Luxembourg, United Arab Emirates, Singapore, Kuwait, the United States. Canada made the 6th position. To avoid implying causation where the situation may be more linked to corelation, it is important to highlight that energy production or consumption is not always the cause of CO_2 emission. Thus, in higlighting trends, a key pointis that 2 countries, Singapore and Luxemborg, that were listed among the top 5 lowest energy producers are also among the top 5 CO_2 emitters. Again, recall that this indicator is also a funtion of population and consumption of CO_2. Also, the United Arab Emirates, a top CO_2 emitter, is among the top 5 energy consumers. Low CO_2 emitters include Zimbabwe, Peru, Indonesia, Colombia, and Vietnam.

Learning note:

- Conclusions is a very important phase of a project especially analysis kind of project wherein you will give your overall intention and interpretation about your analysis. Clarifying the limitations of the study analysis part in the dataset allows the reader to better understand under which conditions the

  results should be interpreted. Clear descriptions of limitations of a study also show that the researcher has a holistic understanding of his/her study. And this is something very positive!. In other words, clearly describing the limitations of your study strengthens your work!.

Communication

Reasoning is provided for each analysis decision, plot, and statistical summary.

Required: ❗

- Please, note that each plot (or set of plots) must be followed by an explanatory paragraph. Please write a few lines after every set of plots describing your observations so that the audience can validate their understanding with your analysis throughout the reading of the project, although later you include these comments in a brief summary in the conclusions section.

  Currently, the set of plots regarding: **"The lowest 5 energy consumers across the years", "The top 5 Consumption CO2 per capita countries" and "The lowest 5 Consumption CO2 per capita countries"** have missing comments describing the observations.

Visualizations made in the project depict the data in an appropriate manner that allows plots to be readily interpreted.

Awesome:

- I have to congratulate you for the graphics that you have included in the project. All the plots have the titles, legends and axes well labeled, making them very easy to understand. **Nice work!**.

Learning note:

- I've seen that you are using Matplotlib to create your visuals. It is very powerful and it is very attractive, which is an important aspect to catch the attention of the reader. I'd like to share with you this very good guide about some advanced visualizations of Matplotlib and Seaborn.

RESUBMIT

DOWNLOAD PROJECT

DOWNLOAD PROJECT

Best practices for your project resubmission

Ben shares 5 helpful tips to get you through revising and resubmitting your project.

Watch Video (3:01)

RETURN TO PATH

Rate this review

START