UDACITY

Return to Classroom

Investigate a Dataset

REVIEW

HISTORY

Meets Specifications

Dear student,
You have done a good job with this investigation task using pandas.
I really love the notebook, it is outstanding and covers beyond what is required, well done

Useful Resources:

- I would like to share this website with your where you can learn about how to deal with missing data. This will be an interesting read:
  https://stefvanbuuren.name/fimd/sec-MCAR.html
- Also, here is an awesome Pandas cheatsheet with you:
  https://pandas.pydata.org/Pandas_Cheat_Sheet.pdf
- This will help you in choosing the right Pandas methods when you do wrangling and exploration.
  This is also a very useful guide on which plot type to use for a specific analysis scenarios.
  http://www.mymarketresearchmethods.com/wp-content/uploads/2013/01/Chart-types.jpg

Code Functionality

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

Awesome job meeting requirement here.

✓Code is run all the lines of the python code without any errors
✓The code produces what is required

Markdown in Jupiter

Try adding markdown as it is an excellent way to have a clear notebook in Jupiter, I use it a lot at work as it reminds me of the business requirement or logic I used especially for issues I run once a year or quarter: https://towardsdatascience.com/jupyter-and-markdown-cbc1f0ea6406

Useful Links

- Most used Pandas functions: https://medium.com/analytics-vidhya/top-20-pandas-functions-which-are-commonly-used-for-exploratory-data-analysis-3cb817a60f46
- Python: https://towardsdatascience.com/tips-and-tricks-for-fast-data-analysis-in-python-f108ad32fa90
- Here is an excellent guide for markdown: https://medium.com/analytics-vidhya/the-ultimate-markdown-guide-for-jupyter-notebook-d5e5abf728fd

- 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 job using the below and vectors instead of lists and dictionaries

- Pandas
- Numpy
- Matplotlib

Why Vectors instead of loops

Vectors and built in functions makes data investigation and analysis fast and accurate, here is a nice article: https://medium.com/analytics-vidhya/understanding-vectorization-in-numpy-and-pandas-188b6ebc5398

Useful Links

Some important Pandas built-in functions:

- Value-Counts
- Indexing and Selecting data
- Group-by

- The code makes use of at least 1 function to avoid repetitive code.
- The code contains good comments and meaningful variable names, making it easy to read.

Awesome job defining and implementing a function that can be used repetitively:

```
#this function will calculate percentage differences for all the health conditions
def calculate_percentage_values(Arrived_with,People_with,Arrived_without,People_without):
    '''
    This Function will help me calculate percentage value difference of those who arrived and those who did not arrive with
    that medical condition
    '''
    a = (Arrived_with/People_with ) * 100
    b = (Arrived_without/People_without) * 100
    print('Percentage of those who arrived with this health condition : {}'.format(a))
    print('Percentage of those who arrived without this health condition : {}'.format(b))
    result = print('The percentage difference for this health condition = {}'.format(a-b))
    return result
```

Suggestion

When writing a function, it is recommended that you explain what the function does not with comments preceding the function definition but with what we call 'docstrings' which are multi-line comments we usually add after the function header.
Check here the importance of commenting your code: https://realpython.com/lessons/importance-writing-good-code-comments/

Quality of Analysis

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

Awesome job done here adding more than is required in clear and explanatory questions:

Based on the data present, I was able to come up with the questions below as to why the patients may not show up for their scheduled medical appointments in Brazil

- Can the age of the patients be a factor for them to show up or not show up for the medical appointments?
- Does the patient's health condition affect if they will show up for their medical appointments?
- Does sending an SMS regarding the medical appointment to the patients help them show up for their appointments?

Suggested links

I will suggest you check the below article showing how to develop analytical questions:
https://www.datapine.com/blog/data-analysis-questions/

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 done here

Awesome done here

✓Well done justifying the reason for certain data cleaning choices using markdown cells explaining the rationale behind those cleaning decisions.

✓Good work in implementing a Data Wrangling Phase!

✓You captured the issues in this dataset and also explained every step and cleaning! Good job!

Useful Links

- As you have learned, missing data is one of the major issues data analysts encounter when they start the exploration. I really recommend you check out this website to learn more details about the different types of missing data and how to deal with each one: https://stefvanbuuren.name/fimd/sec-MCAR.html
- Real Python gives a good overview of examining and cleaning your data

Exploration Phase

- The project investigates the stated question(s) from multiple angles.
- The project explores at least three variables in relation to the primary question. This can be an exploratory relationship between three variables of interest, or looking at how two independent variables relate to a single dependent variable of interest.
- The project performs both single-variable (1d) and multiple-variable (2d) explorations.

Awesome job here.

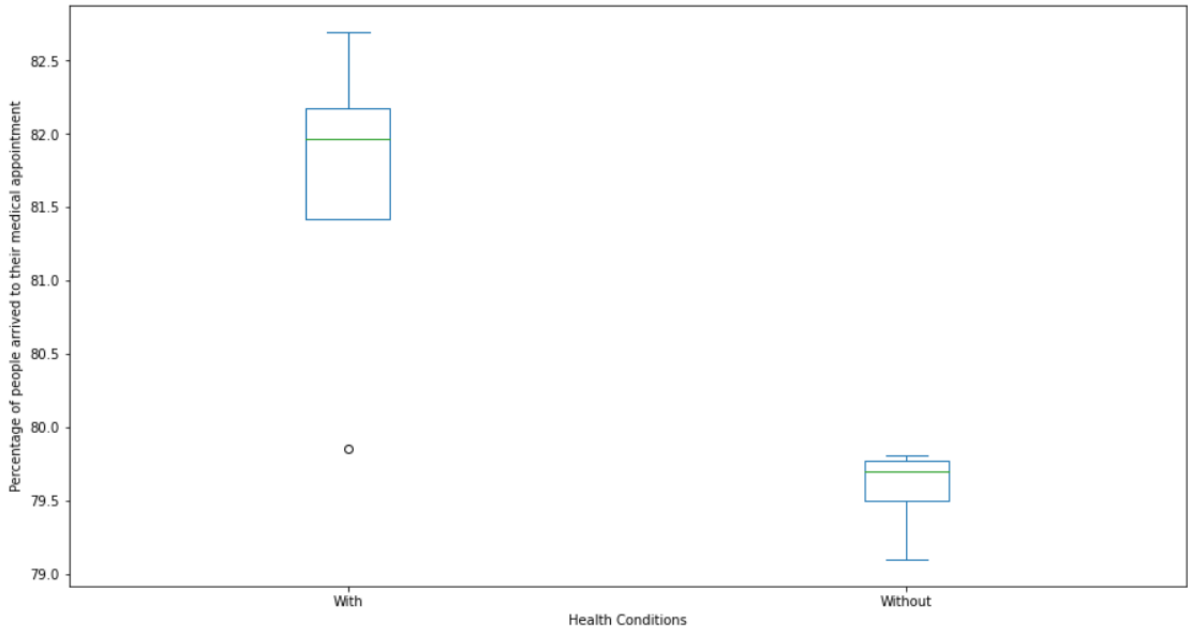
✓ The questions were thoroughly investigated from various angles

✓Used both 1d and 2d explorations were used for several variables investigated..!! Well done..!!

✓Awesome job exploring at least three variables in relation to the primary question

Excellent 1d exploration

```
With = [81.99672667757774,82.69724770642202,81.9400983459991,79.85119047619048]
Without = [79.63890892784028,79.09852233405846,79.76432344575376,79.8070193447243]
index = ['Diabetes', 'Hypertension', 'Handicap', 'Alcoholism']
df_health_conditions = pd.DataFrame({'With': With,'Without': Without}, index=index)
ax = df_health_conditions.plot.box(rot=0,figsize = (15,8))
plt.xlabel('Health Conditions')
plt.ylabel('Percentage of people arrived to their medical appointment')
plt.show()
```



Here are the differences between bivariate and univariate data:

Summary: Differences between univariate and bivariate data.

Univariate Data	Bivariate Data
• involving a **single variable**	• involving **two variables**
• does not deal with causes or relationships	• deals with causes or relationships
• the major purpose of univariate analysis is to describe	• the major purpose of bivariate analysis is to explain
• central tendency - mean, mode, median • dispersion - range, variance, max, min, quartiles, standard deviation. • frequency distributions • bar graph, histogram, pie chart, line graph, box-and-whisker plot	• analysis of two variables simultaneously • correlations • comparisons, relationships, causes, explanations • tables where one variable is contingent on the values of the other variable. • independent and dependent variables

Sample question: How many of the students in the freshman class are female?	**Sample question:** Is there a relationship between the number of females in Computer Programming and their scores in Mathematics?

Suggested Reads

Useful link on how to best choose the plot types: https://udacity-reviews-uploads.s3.us-west-2.amazonaws.com/_attachments/83662/1588660779/How-to-Visualize-your-Data-with-Charts-and-Graphs.jpg

- The project's visualizations are varied and show multiple comparisons and trends.
- At least two kinds of plots should be created as part of the explorations.
- Relevant statistics are computed throughout the analysis when an inference is made about the data.

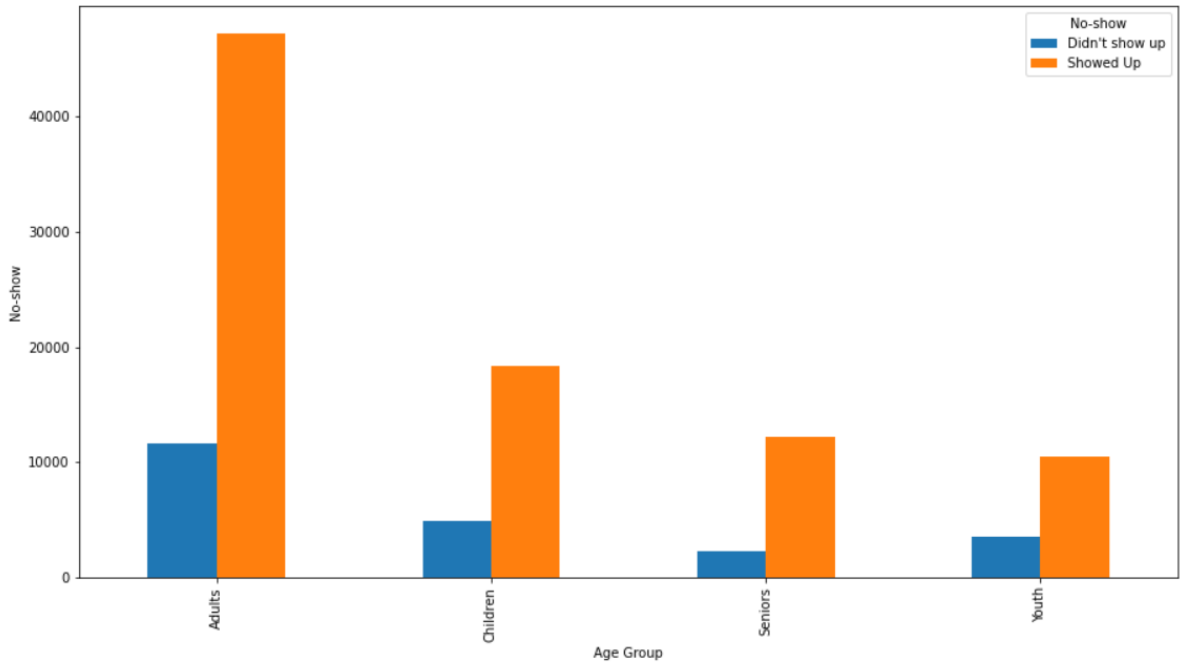
Well done meeting requirement here for using more than one plot type:

✓ You successfully used box plots
✓ Well implemented bar plots
✓ Well done having more than 2 plots to answer questions

Outstanding

I love that you used all the attributes of the .plot() function in Matplotlib:

```
#Plotting a bar graph to get some analysis
df = df_appointment.groupby(['Age_group','No-show']).size()
df=df.unstack()
df.plot(kind='bar', figsize = (15,8))
plt.xlabel('Age Group')
plt.ylabel('No-show')
plt.show()
```



Suggested Galleries

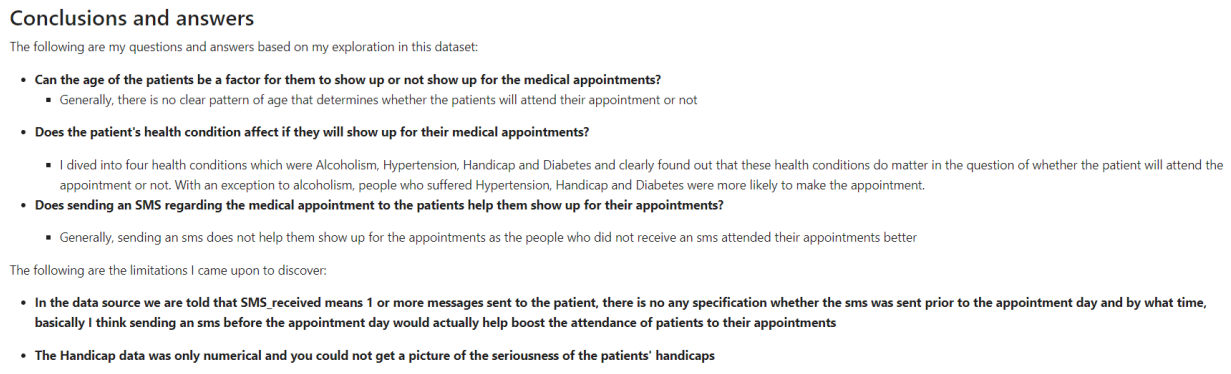
- The Python Graph Gallery contains a gallery of different visualisation and template code you can use for your visualisations.
- Another gallery I'd recommend you is the seaborn gallery

Conclusions Phase

- The Conclusions have reflected on the steps taken during the data exploration.
- The Conclusions have summarized the main findings in relation to the question(s) provided at the beginning of the analysis accurately.
- The project has pointed out where additional research can be done or where additional information could be useful.
- The conclusion should have at least 1 limitation explained clearly.
- The analysis does not state or imply that one change causes another based solely on a correlation.

Awesome

✓ You have taken your time in thinking and presenting a brief but very clear conclusion about your findings. This is a very important part of the report since many readers have the practice of going through the conclusion before reading the analysis section that lead to it.
✓Excellent listing of clear limitations in the dataset



Conclusions and answers

The following are my questions and answers based on my exploration in this dataset:

- **Can the age of the patients be a factor for them to show up or not show up for the medical appointments?**
  - Generally, there is no clear pattern of age that determines whether the patients will attend their appointment or not
- **Does the patient's health condition affect if they will show up for their medical appointments?**
  - I dived into four health conditions which were Alcoholism, Hypertension, Handicap and Diabetes and clearly found out that these health conditions do matter in the question of whether the patient will attend the appointment or not. With an exception to alcoholism, people who suffered Hypertension, Handicap and Diabetes were more likely to make the appointment.
- **Does sending an SMS regarding the medical appointment to the patients help them show up for their appointments?**
  - Generally, sending an sms does not help them show up for the appointments as the people who did not receive an sms attended their appointments better

The following are the limitations I came upon to discover:

- **In the data source we are told that SMS_received means 1 or more messages sent to the patient, there is no any specification whether the sms was sent prior to the appointment day and by what time, basically I think sending an sms before the appointment day would actually help boost the attendance of patients to their appointments**
- **The Handicap data was only numerical and you could not get a picture of the seriousness of the patients' handicaps**

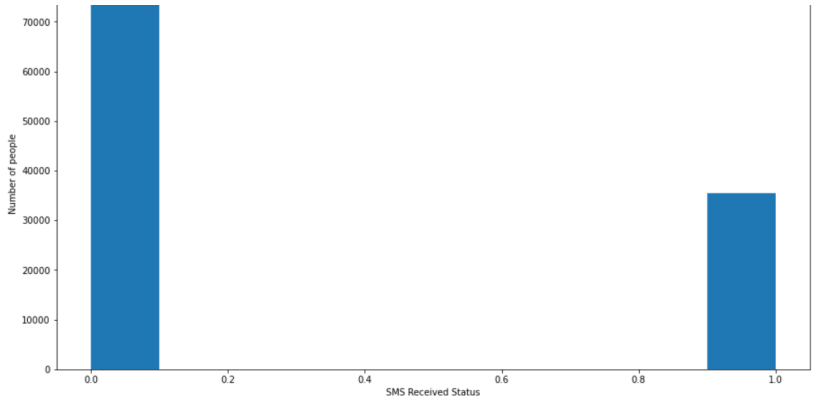
Communication

- The code should have ideally the following sections: Introduction; Questions; Data Wrangling; Exploratory Data Analysis; Conclusions, Limitation.
- Reasoning is provided for each analysis decision, plot, and statistical summary.
- Interpretation of plots and application of statistical tests should be correct and without error.

- Comments are used within the code cells.
- Documented the flow of analysis in the mark-down cells.

Awesome job organizing your EDA notebook properly.

✓Reasoning is provided for each analysis decision, plot, and statistical summary.
✓Comments are used within the code cells.
✓Documented the flow of analysis in the mark-down cells.



From the previous argument, here we can also see that People who received an sms(1) are fewer than those who did not receive an sms(0) and actually in that state, many of this fewer number did not attend their appointments thus sending an sms did not bring about any effect to the attendance of the patients to their medical appointments

Useful Links

Here is an excellent guide for markdown: https://medium.com/analytics-vidhya/the-ultimate-markdown-guide-for-jupyter-notebook-d5e5abf728fd

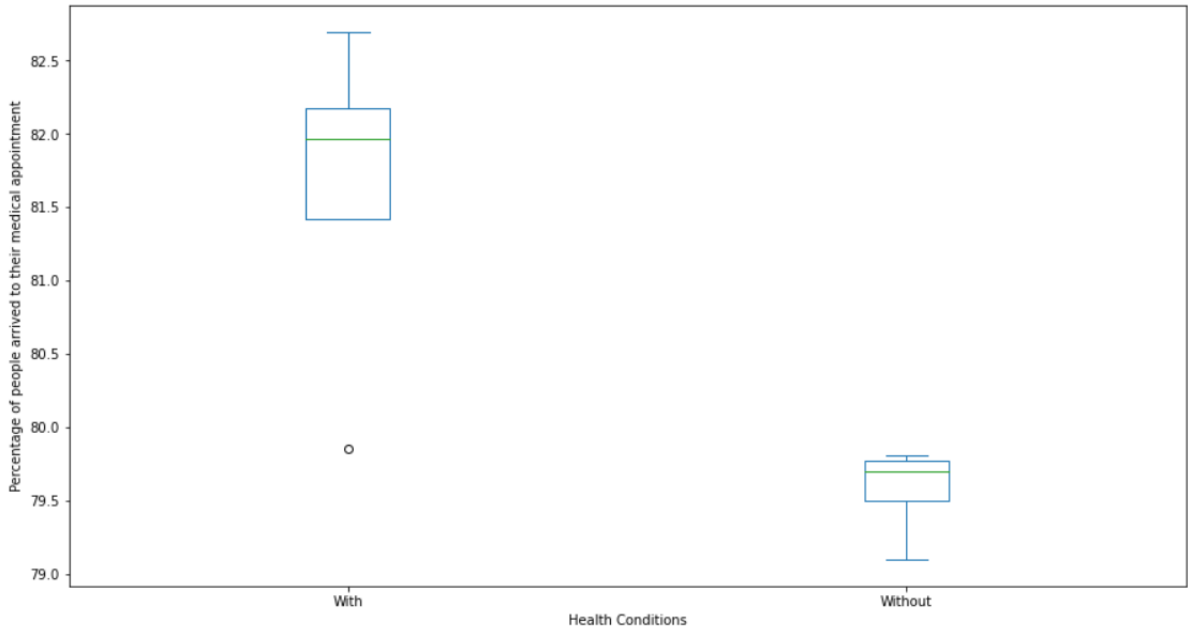
Visualizations made in the project depict the data in an appropriate manner (i.e., has appropriate labels, scale, legends, and plot type) that allows plots to be readily interpreted.

Excellent job having:
✓Clear abbreviations labels on all plots
✓Plot titles
✓Legends visible

Outstanding

In each plot you have all components (label, title and legends) as shown below:

```
With = [81.99672667757774,82.69724770642202,81.9400983459991,79.85119047619048]
Without = [79.63890892784028,79.09852233405846,79.76432344575376,79.8070193447243]
index = ['Diabetes', 'Hypertension', 'Handicap', 'Alcoholism']
df_health_conditions = pd.DataFrame({'With': With,'Without': Without}, index=index)
ax = df_health_conditions.plot.box(rot=0,figsize = (15,8))
plt.xlabel('Health Conditions')
plt.ylabel('Percentage of people arrived to their medical appointment')
plt.show()
```



Useful Links

Here is a nice document on the importance of labels: https://teach.files.bbci.co.uk/skillswise/ma37grap-e3-f-using-clear-labels-on-your-chart-or-diagram.pdf

DOWNLOAD PROJECT

RETURN TO PATH

Rate this review

START