Write a Data Science Blog Post

REVIEW | CODE REVIEW | HISTORY

Meets Specifications

Dear Student,
Excellent job submitting a great project 👍
And congratulations you did very well.

Kindly go through the suggestions carefully and please take them positively and constructively as an opportunity to learn and grow and also improve the overall quality of your work. Learning in data science never stops and as you will see, there is not always one perfect way to do things but intuition and experimentation are very often required depending on the situation.

For further learning in clustering and as a general companion to the Nanodegree, I recommend the classic Harvard CS109 course here - http://cs109.github.io/2015/pages/videos.html

I also recommend studying good Kaggle kernels like this one - https://www.kaggle.com/arthurtok/principal-component-analysis-with-kmeans-visuals?scriptVersionId=1543947

and good blogs like this one - https://machinelearningmastery.com/?s=clustering&post_type=post&submit=Search

Also If this review helped you in some way, Please consider rating it positively.
I wish you all the best for your future endeavors. Keep learning and stay udacious!

Code Functionality and Readability

✓ **All the project code is contained in a Jupyter notebook, which demonstrates successful execution and output of the code.**

✓ **Code has easy-to-follow logical structure. The code uses comments effectively and/or Notebook Markdown cells correctly. The steps of the data science process (gather, assess, clean, analyze, model, visualize) are clearly identified with comments or Markdown cells, as well. The naming for variables and functions should be according to PEP8 style guide.**

Good job with this one!
The Code has easy-to-follow logical structure. Nice work using comments in the code effectively and I noticed the steps of the data science process were all followed correctly. Good job too with the naming convention. :+1:

Consider going through the links below to gain more insights on code readability improvements.
Python Coding Style & Standards - http://www.voidspace.org.uk/python/articles/python_style_guide.shtml

✓ **Code is well documented and uses functions and classes as necessary. All functions include document strings. DRY principles are implemented.**

Quite nice work.
It's nice to have document strings in all your functions so that it is clear to anyone looking at the function as to what it does and what are the parameters or what the function returns.

For document strings you might also check out the following -
This post regarding Docstrings vs Comments : https://stackoverflow.com/questions/19074745/docstrings-vs-comments
Google Style Python Docstrings : https://sphinxcontrib-napoleon.readthedocs.io/en/latest/example_google.html

Data

✓ **Project follows the CRISP-DM process outlined for questions through communication. This can be done in the README or the notebook. If a question does not require machine learning, descriptive or inferential statistics should be used to create a compelling answer to a particular question.**

CRISDM could be more clearer and explicit in the notebook.
A good idea is to have each of a structure from each of the step of CRISPDM and in each structure the corresponding code and markdown should be incorporated.

Please refer these slides with the highlighted points for each step from CRISPDM for more details - https://paginas.fe.up.pt/~ec/files_0405/slides/02%20CRISP.pdf

It might be good to consider restructuring code with each of the parts of CRISP-DM directly stated with following markdown and code. This is a time consuming portion of writing code (restructuring), but it can be super useful for your readers.

You should in particular have these sections in your notebook

- Business understanding: outline the questions you will answer along with why they are relevant and important
- Data understanding: provide some stats about your data like mean and std deviations of the different features
- Data preparation: show how you clean and prepare the data. It is important to document and not just write the code
- Data modeling: if you created a model this is the section you would put it in or the analysis
- Results evaluation: conclusion and discussion

✓ **Categorical variables are handled appropriately for machine learning models (if models are created). Missing values are also handled appropriately for both descriptive and ML techniques. Document why a particular approach was used, and why it was appropriate for a particular situation.**

Nice job documenting your process for categorical variables. Might want to also even look into using likelihood encoding of categorical features(https://www.kaggle.com/tnarik/likelihood-encoding-of-categorical-features) or smoothing (https://www.kaggle.com/ogrellier/python-target-encoding-for-categorical-features).

And amazing job with missing values as well especially for question 1, flawless work for that one.
Although I felt that for the reasoning for dropping the 75% rows from salary column from `activity_df` could be justified in a better way, but nevertheless good job.

Analysis, Modeling, Visualization

✓ **There are between 3-5 questions asked, related to the business or real-world context of the data. Each question is answered with an appropriate visualization, table, or statistic.**

Github Repository

✓ **Student must have a Github repository of their project. The repository must have a README.md file that communicates the libraries used, the motivation for the project, the files in the repository with a small description of each, a summary of the results of the analysis, and necessary acknowledgements. Students should not use another student's code to complete the project, but they may use other references on the web including StackOverflow and Kaggle to complete the project.**

Overall repository is well maintained.

Some other suggestions which may not be required from rubrics but just for improvement:
1) Considering adding a separate licence file other than a section in READMe which specifies the how your code can be re-used.
You can refer to this site - https://choosealicense.com/
or create your own kind.

2) Remove un-required files from your repo such as `.DS_Store` (A Mac OS generated file) and `.ipynb_checkpoints` .
You can do manually remove them or Consider adding a `.gitignore` file which will ensure that git ignores unecessary files while you commit.
Refer here for more details - https://git-scm.com/docs/gitignore
And how to create your own gitignore from here - https://www.gitignore.io/

3) Git commits.
The commits you've made to the repo are good. Adopting to a standard good practice would be beneficial in the long run.
Refer this guide for more details - https://github.com/trein/dev-best-practices/wiki/Git-Commit-Best-Practices

Blog Post

✓ **Student must have a blog post on a platform of their own choice (can be on their website, a Medium post or Github blog post). Student must communicate their results clearly. The post should not dive into technical details or difficulties of the analysis - this should be saved for Github. The post should be understandable for non-technical people from many fields.**

Well written Blog !
The results are communicated clearly. The post is well balanced not too many technical details or difficulties of the analysis.
The post is well understandable, even for non-technical people from many fields.
Great job !

✓ **Student must have a title and image to draw readers to their post.**

Nice images and Title.
Please consider attributing the images to the source In case you are not the owner of them. This is considered very professional to credit the right sources.

✓ **There are no long, ongoing blocks of text without line breaks or images for separation anywhere in the post.**

Your blog post is very easy to follow. Nice use of short and concise topics for the reader.

You might also consider using some italicizes or bold text (not just titles) in your medium post. Really helps the reader determine where to focus on.

✓ **Each question is answered with a clear visual, table, or statistic that provides how the data supports or disagrees with some hypothesis that could be formed by each question of interest.**

Each question is answered with a relevant visual answering the question. Here are 25 tips that might help you in your future visualizations: https://www.columnfivemedia.com/25-tips-to-upgrade-your-data-visualization-design

DOWNLOAD PROJECT

RETURN TO PATH

Rate this review