**DSPT10 - Portfolio Project**

The Unit 1 Portfolio Project provides you with a first experience selecting a dataset, determining statistical hypotheses, drawing conclusions and presenting your results.

Your final Portfolio Project will be a blog post or post on Medium containing the following sections:

1. Introduction

* Provide relevant background information on your topic.

1. Description of the data

* Provide a description of your dataset. Where did you find it? What data does it contain? Are there any limitations to your dataset?

1. Statistical methods

* Describe your hypotheses.
* Describe any data wrangling or feature engineering you had to do.
* Provide a description of the statistical methods you used to analyze your data.

1. Results

* Describe the results of your analyses.
* Present two visualizations related to your hypotheses.

1. Conclusion

* Report the conclusions of your hypothesis test(s).
* Summarize the key features of your visualizations.
* Discuss any limitations of your analyses.
* Compare your results to other, similar analyses.
* Discuss new questions your results raise or new directions for further research.

You will also give a 2-minute presentation of your results for your cohort during the Build Week (the last week of Unit 1, that’s Sprint 4, week B).

Although you will present your Portfolio Project during the Unit 1 Build Week, you will complete the steps of the Portfolio Project during Sprints 3 - 4. To help you manage your time, we suggest accomplishing the following milestones.

**During Sprint 3, Week B:**

1. If you did not watch the previous cohort present their presentations, watch the video recording. <https://youtu.be/iL9Plal1VLU>
2. Begin looking for interesting datasets. See the end of this document for a list of places to search.
3. Continue looking at datasets. Begin to download and manipulate one or two that look interesting. Start to work on any data wrangling and/or feature engineering you need to do.
4. Sketch out your hypotheses and do background research.
5. Begin to work on your analysis and visualizations.

**During Sprint 4, Week A:**

1. Continue working/refining your analyses and visualizations.
2. Conduct additional background research, and begin writing your blog/Medium post.
3. Video on how to set up Github web pages: [Watch this video for more information](https://www.youtube.com/watch?v=gxbNCGERUjc&feature=youtu.be&ab_channel=LambdaSchool).
4. Continue writing your blog/Medium post.
5. Ask for feedback from peers and make any necessary revisions.
6. Practice your 2-minute talk.

Data sources:

Pretty much every major US governmental agency (CDC, Bureau of Labor Statistics, The Census, etc.) maintains a treasure trove of data.

There's also:

Datasets from the University of Florida Department of Statistics: <http://users.stat.ufl.edu/~winner/datasets.html>

Pew Research: <https://www.pewresearch.org/>

General Social Survey: <https://gss.norc.org/>

Ancestry.com Free data collections: <https://www.ancestry.com/search/categories/freeindexacom/?clickref=1101lbhdWxys&key=Uhttps%3A%2F%2Fwww.ancestry.com%2Fsearch%2Fcategories%2Ffreeindexacom%2F%3Fclickref%3D1101lbhdWxys&o_lid=01011l4pVw&o_sch=Affiliate+External&o_xid=01011l4pVw>

Popular baby names by year since the founding of the SSA: <https://www.babynamewizard.com/>

<https://data.world/>

<https://www.kaggle.com/datasets>

<https://catalog.data.gov/dataset>

<https://datasetsearch.research.google.com/>

<https://archive.ics.uci.edu/ml/index.php>

<https://github.com/plotly/datasets>

<https://lionbridge.ai/datasets/20-best-image-datasets-for-computer-vision/>

<https://www.visualdata.io/>

<https://pathmind.com/wiki/open-datasets>

<https://blog.playment.io/top-open-image-dataset/>

<https://blog.cambridgespark.com/>

<https://www.analyticsvidhya.com/blog/2018/03/>

<https://data-flair.training/blogs/machine-learning-datasets/>