

## Structures of my project

OneDrive - Personal > Documents > nanodegree > DAND > 5\_Data\_Visualization > 8\_Project > \_submit\_project > wrangle >

Name	Status	Date modified	Type	Size
<b>A</b> raw_data	✓	21/01/2019 19:44	File folder	
<b>B</b> wrangling	✓	21/01/2019 19:44	File folder	
y_2006_data 1	✓	05/01/2019 03:30	File	16,885 KB
y_2007_data 2	✓	05/01/2019 03:33	File	17,607 KB
y_2008_data 3	✓	06/01/2019 09:15	File	16,732 KB
y_2008_january_data 4	✓	09/01/2019 11:10	File	144,387 KB

**A:** This is where I put my raw data. If you click this folder, you will find A1 and A2. More about this later.

**B:** This is where I do all my wrangling from year 2006, 2007, and 2008. More about this later.

**1,2, and 3:** This is where I put my sampld data after wrangling is done in folder B. Each file contains about 70,000 rows. It is about 1% of original raw data.

**4:** This file only contains only January, 2018 data. It is wrangled data.

OneDrive - Personal > Documents > nanodegree > DAND > 5\_Data\_Visualization > 8\_Project > \_submit\_project > wrangle > raw\_data >

Name	Status	Date modified	Type	Size
2006	✓	21/01/2019 19:44	File folder	
2007	✓	21/01/2019 19:44	File folder	
2008	✓	21/01/2019 19:44	File folder	
supplement	✓	21/01/2019 19:44	File folder	
link	✓	13/12/2018 00:26	Text Document	1 KB

**A1:** This is where I put raw data. Each folder contains about 800 Mb of data! Therefore you have to manually download it yourself and put it into these folders: 2006, 2007, and 2008. Here is the link to raw data: <http://stat-computing.org/dataexpo/2009/the-data.html>. Please download only 2006, 2007, and 2008.

**A2:** This folder contains two raw data: airports.csv, and carriers.csv. These are helper files

Project 5: Data Visualization. Dataset: Flights  
By Andy Soelistio. January 21, 2019

that make my data more human-intelligible. Here is the link if you to download it yourself:  
<http://stat-computing.org/dataexpo/2009/supplemental-data.html>

OneDrive - Personal > Documents > nanodegree > DAND > 5_Data_Visualization > 8_Project > _submit_project > wrangle > wrangling					
Name	Status	Date modified	Type	Size	
.ipynb_checkpoints	✓	21/01/2019 19:44	File folder		
wrangle_jan_y_2008.ipynb <b>B1</b>	✓	09/01/2019 13:48	IPYNB File	1,748 KB	
wrangle_y_2006.ipynb <b>B2</b>	✓	12/01/2019 03:11	IPYNB File	1,700 KB	
wrangle_y_2007.ipynb <b>B3</b>	✓	12/01/2019 03:09	IPYNB File	1,764 KB	
wrangle_y_2008.ipynb <b>B4</b>	✓	15/01/2019 05:50	IPYNB File	1,755 KB	

This folder contains files where I do most of my data-wrangling. This is not a trivial task.

**B1:** This is where I wrangle data. The output of this file is saved as 4.y 2008 january data. This file contains 600,000 rows! Just January 2008 data.

**B2,B3,and B4:** This is where I wrangle data. The output of this file is save as 1.y 2006 data, 2.y 2007 data, 3.y 2008 data. Each file contains 70,000 rows of sampled data.

OneDrive - Personal > Documents > nanodegree > DAND > 5_Data_Visualization > 8_Project > _submit_project					
Name	Status	Date modified	Type	Size	
.ipynb_checkpoints	✓	21/01/2019 19:54	File folder		
pictures	✓	21/01/2019 20:23	File folder		
reveal.js	✓	21/01/2019 19:54	File folder		
wrangle <b>W1</b>	✓	21/01/2019 19:44	File folder		
explanatory_presentation.ipynb <b>M1</b>	✓	21/01/2019 20:04	IPYNB File	5,594 KB	
explanatory_presentation.slides <b>M2</b>	✓	21/01/2019 20:08	HTML File	3,497 KB	
exploratory_analysis.ipynb <b>M3</b>	✓	21/01/2019 19:55	IPYNB File	5,586 KB	
output_toggle	✓	12/03/2018 16:08	TPL File	1 KB	
readme	↻	21/01/2019 21:46	Microsoft Word Doc...	139 KB	

This folder is where I put my project files.

**W1:** This is where I put my raw data and wrangling files discussed above.

**M1:** This is my explanatory presentation ipynb file.

**M2:** Click this to view slideshow presentation.

**M3:** This is where I do further wrangling and exploratory analysis.

## Creating slideshow

In case if you need to create slideshow yourself, please type this on your command prompt or bash:

```
andy@OIC MINGW64 ~/OneDrive/Documents/nanodegree/DAND/5_Data_Visualization/8_Project/__submit_project
$ ls
explanatory_presentation.ipynb    exploratory_analysis.ipynb  pictures/    reveal.js/
explanatory_presentation.slides.html  output_toggle.tpl          readme.docx  wrangle/

andy@OIC MINGW64 ~/OneDrive/Documents/nanodegree/DAND/5_Data_Visualization/8_Project/__submit_project
$ jupyter-nbconvert --to slides explanatory_presentation.ipynb --reveal-prefix=reveal.js --template output_toggle
[NbConvertApp] Converting notebook explanatory_presentation.ipynb to slides
[NbConvertApp] Writing 3565522 bytes to explanatory_presentation.slides.html

andy@OIC MINGW64 ~/OneDrive/Documents/nanodegree/DAND/5_Data_Visualization/8_Project/__submit_project
$
```

## Main Findings

Since I have done extensive comment on exploratory analysis (and I'm short on time 😊), I'll just copy and paste from my jupyter notebook:

**Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?**

- From [Section 5.2](#), [Section 5.3](#), and [Section 5.4](#), we can see a general trend. I call it **cascading** delays. Shortest mean delays started in the morning or late morning and then delays get progressively longer at Late Night or Early Morning. This trend makes intuitive sense because if we had delays in the morning, then delays will be carried over in the next flight and cascaded as hours go by.
- I decided to check / drill down this hypothesis -- `that delays are cascaded`. I made 3 queries: [query 1](#), [query 2](#), and [query 3](#). The query results surprised me and I have to reject the hypothesis. 300 minutes delays happened in the first or first two rows and happened in the Early Morning. Then flights generally returned to normalcy, i.e no delays as hours progressed.
- Then I explored causes of delays. I plot the causes of delays and it strengthens the **rejection** of the hypothesis -- `that delays are cascaded`. [From the plot](#), we can see that delays are caused by LateAircraftDelay, CarrierDelay and WeahterDelay that happened longer at LateNight and EarlyMorning.

## Were there any interesting or surprising interactions between features?

- [Fig. 5.7a](#) and [Fig. 5.7b](#) show that big-sized airports have shorter mean delays. Notice that they have less red and yellow circles.
- Almost all plots in [Section 5.7a](#) and [Section 5.7b](#) show that airports located on Western coast have shorter mean delays than airports located on Eastern Coast. Also there are less airports located on the Western coast than the Eastern coast.

## Explanatory Presentation

My explanatory presentation is basically from my exploratory\_analysis.ipynb. I skipped the cells that I don't need. I select cells that I think is important to be slides. Hope you enjoy my presentation.

## References:

Slideshow presentation:

<https://www.youtube.com/watch?v=EOpcxyORA1A&index=148&t=0s&list=WL>

TimeSeries:

<https://towardsdatascience.com/basic-time-series-manipulation-with-pandas-4432afee64ea>

plotly:

<https://medium.com/analytics-vidhya/introduction-to-interactive-geoplots-with-plotly-and-mapbox-9249889358eb>

online tutorials:

nanodegree tutorials