

Introduction to Data Science

Data Visualization and Effective Storytelling

Joanna Bieri DATA101

Important Information

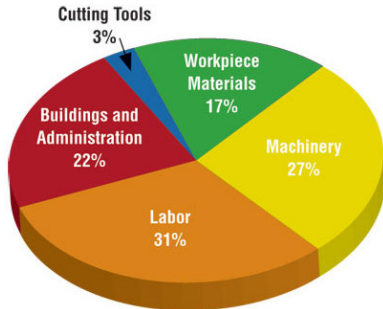
- Email: joanna_bieri@redlands.edu
- Office Hours: Duke 209 [Click Here for Joanna's Schedule](#)

Announcements

In TWO WEEKS - Data Ethics You should have some resources (book or 3-4 articles) about some area of data science ethics/impacts.

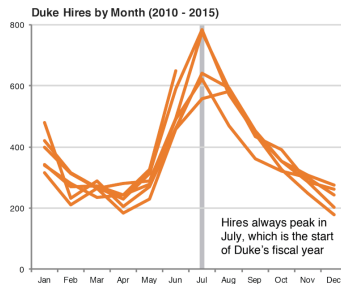
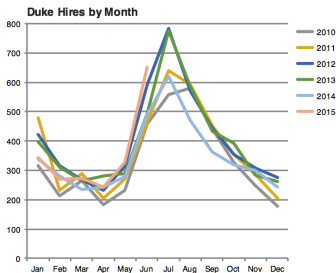
How to make better Data Visualizations

- 1 Keep it simple!
- 2 Make it easy to see the distinction between categories.



How to make better Data Visualizations

- 3 Use color to draw attention
- 4 Tell a story - annotate if needed.



Principles

- ① Order Matters
- ② Put long categories on the y-axis
- ③ Pick a Purpose.
- ④ Keep scales consistent
- ⑤ Select meaningful colors
- ⑥ Use meaningful and nonredundant labels.

Let's work through an example.

Data - September 2019 - YouGov survey asked 1,639 adults in Great Britain:

In hindsight, do you think Britain was right/wrong to vote to leave EU?

They had the choices:

- Right to leave
- Wrong to leave
- Don't know

Source: YouGov Survey Results, retrieved Oct 7, 2019

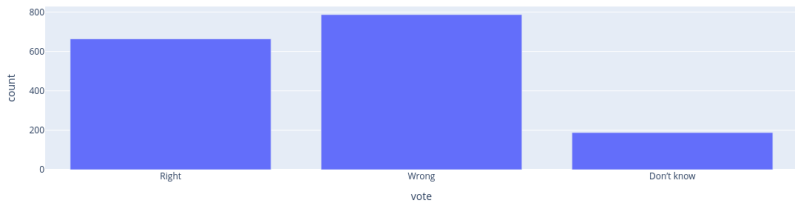
Here is the data

	vote	location	count
0	Right	total	664
1	Wrong	total	787
2	Don't know	total	188
3	Right	london	63
4	Wrong	london	110
5	Don't know	london	24
6	Right	rest_of_south	241
7	Wrong	rest_of_south	257
8	Don't know	rest_of_south	49
9	Right	midlands_wales	145
10	Wrong	midlands_wales	152
11	Don't know	midlands_wales	57
12	Right	north	176
13	Wrong	north	176
14	Don't know	north	48

Basic bar plot - all votes

```
mask = DF['location'] == 'total'  
DF_plot=DF[mask]  
  
fig = px.bar(DF_plot,x='vote',y='count')  
  
fig.show()
```

Basic bar plot - all votes



Order Matters! Reorder the data by frequency.

Sometimes we want the labels to be alphabetical.

```
yaxis={'categoryorder': 'category descending'}
```

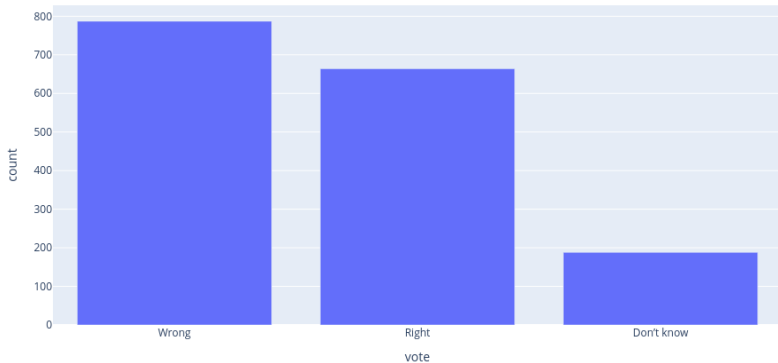
Here are all the possible options for the 'categoryorder':

```
['trace', 'category ascending', 'category descending',  
'array', 'total ascending', 'total descending', 'min  
ascending', 'min descending', 'max ascending', 'max  
descending', 'sum ascending', 'sum descending', 'mean  
ascending', 'mean descending', 'median ascending', 'median  
descending']
```

Order Matters! Reorder the data by frequency.

```
fig = px.bar(DF_plot,x='vote',y='count')  
  
fig.update_layout(xaxis={'categoryorder': 'total descending'},  
                  autosize=False,  
                  width=800,  
                  height=500)  
  
fig.show()
```

New plot

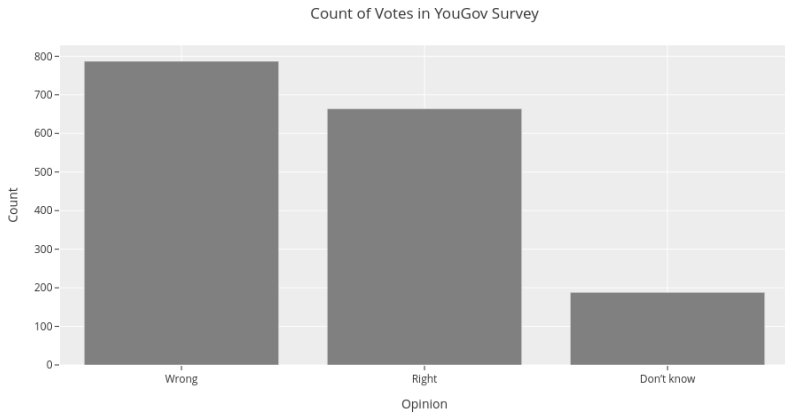


Fix up the figure options

How do we:

- Add a title
- Add x and y axis labels
- Resize
- Choose gray color
- use ggplot2 template

New Plot



Basic Bar Plot - votes by region

First we need to get some data about the votes per region.

```
DF_plot = DF.groupby('location',as_index=False).sum().copy()
```

	location	vote	count
0	london	RightWrongDon't know	197
1	midlands_wales	RightWrongDon't know	354
2	north	RightWrongDon't know	400
3	rest_of_south	RightWrongDon't know	547
4	scot	RightWrongDon't know	141
5	total	RightWrongDon't know	1639

Basic Bar Plot - votes by region



Choose a better order of the categories

In the original survey the order was presented as:

```
['london', 'rest_of_south', 'midlands_wales', 'north', 'scot']
```

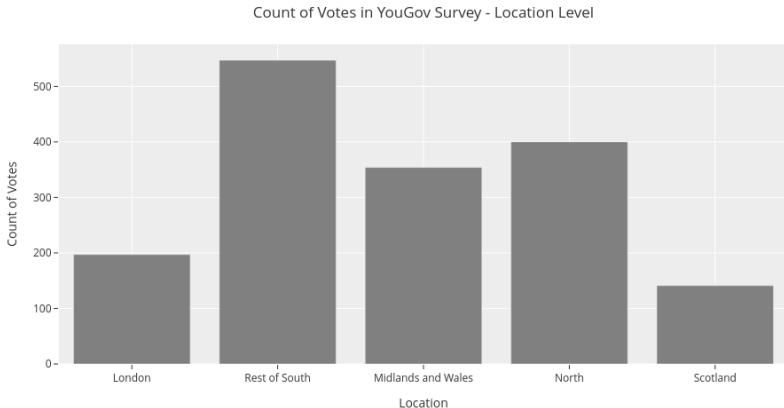
We can choose a manual ordering using

```
xaxis={'categoryorder': 'array',  
      'categoryarray': ['london',  
                        'rest_of_south',  
                        'midlands_wales',  
                        'north',  
                        'scot']}
```

Choose a better order of the categories

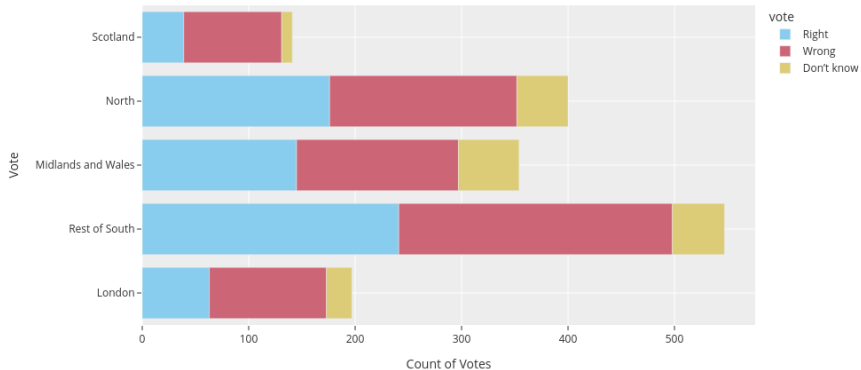
Here I also fixed the label names!

```
DF_plot['location']=  
    DF_plot['location'].replace(old_category,new_category)
```

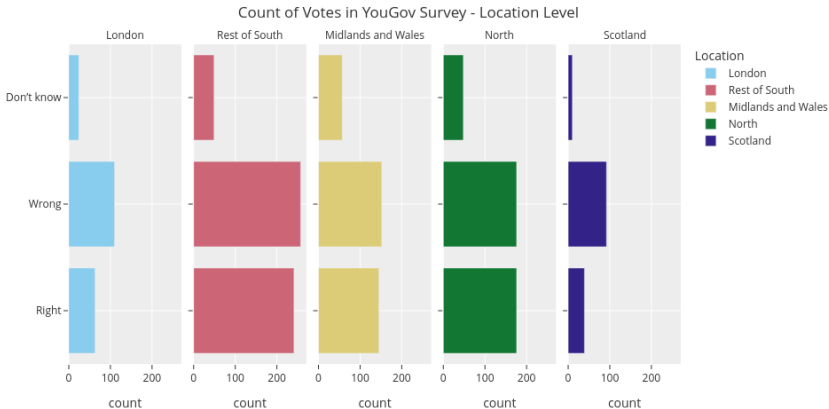


New Bar Plot

Count of Votes in YouGov Survey - Location Level

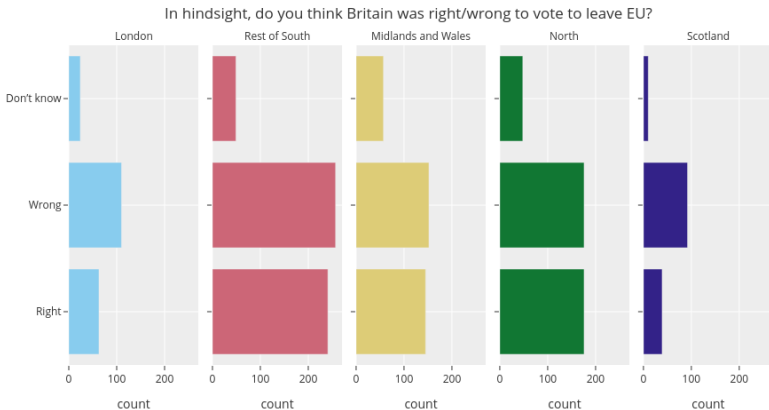


New Bar Plot - Facets



Avoid Redundancy

Here is the same graph again, but avoiding redundancy.



Selecting meaningful colors.

Here is a great website for selecting beautiful colors!

colorbrewer.org

Here is code with new colors

```
fig = px.bar(DF_plot,y='vote',x='count',  
             color='vote',  
             facet_col='location',  
             color_discrete_map={'Right': '#91bfdb',  
                                 'Wrong': '#fc8d59',  
                                 "Don't know": '#ffffbf'})
```


Here is code for better labels

```
fig.for_each_annotation(lambda a:  
    a.update(text=a.text.split("=")[1]))
```

Here is code with new colors

```
my_categories = ['Dont know', 'Right', 'Wrong']
fig.update_layout(yaxis={'categoryorder': 'array',
                        'categoryarray': my_categories},
                  title='In hindsight, do you think
                        Britain was right/wrong
                        to vote to leave EU?',
                  yaxis_title="",
                  template='ggplot2',
                  legend_title='Location',
                  autosize=False,
                  width=1000,
                  height=500,
                  showlegend=False)

fig.show()
```

The Final Plot

