**Report on Efforts of Data Analysis and Visualization**

**On the Co2 Emissions Dataset**

# Getting Started

Provided that the datasets were cleaned and needed no further cleaning, the three datasets were combined together using the common primary keys in each dataset. This step was done after refining and renaming the matching keys together. After that, the unnecessary columns in the datasets were dropped.

# Getting Global Insights

In order to understand the dataset, we need to fetch some information like:

* Lord is the highest user in co2 emissions and sum of trip duration
* Jon was the top user in the distance traveled
* France is the highest country in co2 emissions
* Europe is the highest continent in co2 emissions
* 2015 was the highest year in co2 emissions
* June was the highest month in co2 emissions

# Focusing on Subsets

Now to get a deeper analytical view for the dataset, I went further and found the following interesting facts:

* Reach reported the highest regions of co2 emissions.
* Faith of the Seven was far onto the top in co2 emissions.
* Blackwood became first in co2 emissions.
* Trains was reported to be used 1442 times.
* Lord (highest user in co2 emissions) has done 363 trips.
* The busiest days of the month where 1 and 14 and 25. (Beginning, Half, End)

# Comparing Subsets

Now that we have a deeper understanding to the data, it is time to make some comparison to get actionable insights.

For instance, **we know from General insights that France is the highest in carbon emissions, but what are the highest houses in France?**

* **Blackwood is the highest house in France in co2 emissions.**

**Now, we need to define the relationship between the duration of trip and the CO2 emissions**

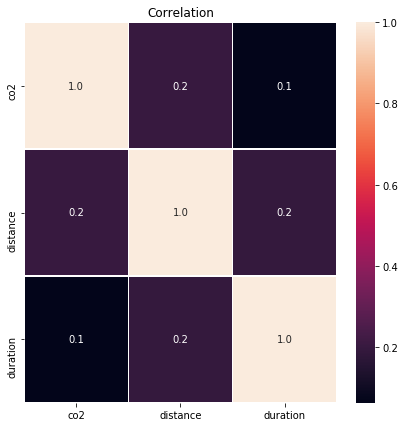
* The correlation coefficient between duration of the trip and the CO2 is small, thus we conclude that co2 emissions is not dependent or linearly related to the trip duration.

**We know from the global insights that France comes first in co2 emissions and then comes the United States. But does it have something to do with the year? Below is the highest 3 years of emissions only in France and only in United States (Each one Separately)**

* Now, it is pretty clear that 2015 and 2016 made it to the top 3 years in both countries, which points out that more activity in these two years was clearly an indicator in having more co2 emissions, in two different countries, not depending on the place.

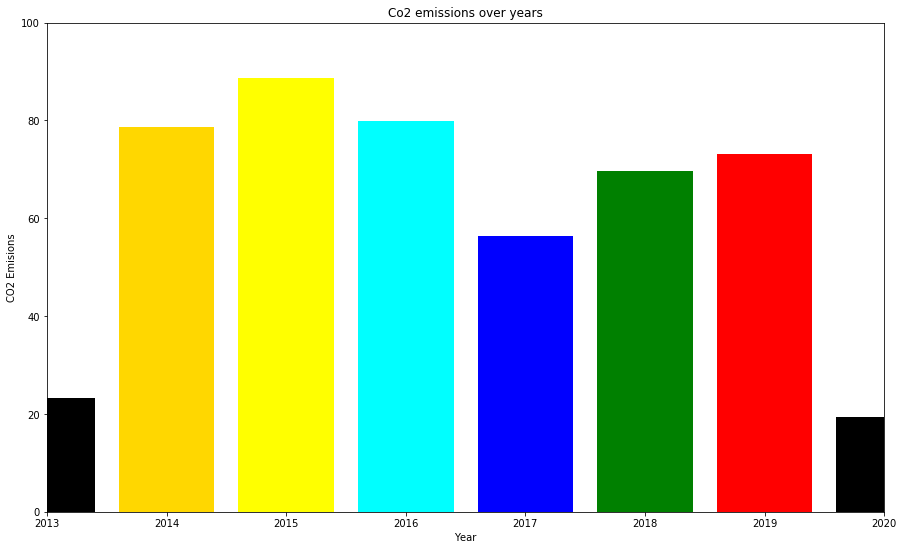
# Visualization

**Now, we need to make a correlation map between the Trip Duration, Distance, and Co2 emissions to be able to detect a clear linear relationship that could spot the light on the main reason of co2 increased emissions.**

****

Now, it is clear that it is not the distance, nor the duration imply the increased emissions of the co2 with a low correlation coefficient of 0.1 and 0.2.

**However, to make further analysis to clarify the relationship between co2 emissions and time**

****

It is clear that in 2014 co2 emissions had a huge spike and continuing to increase through 2015, but since then it has been decreasing. That could be due to shifting into alternative energy sources and country regulations.

Finally, interactive visualizations were provided to allow the user to interact and select, for example, country, month, day, or a year and get a customized graph.