Project Potential Idea

For this project, I am planning to analyze the taxi data. I am planning to either use Uber Data available on Kaggle or the NYC Yellow Cab data from NYC open data to do the analysis. I more inclined towards Yellow Cab data as the Uber data is extremely large and my laptops takes a really long time to work with it, but I am still considering my options. The first part of project will involve in-depth analysis of this data using ggplot, lubridate, tidyr, scales, joins to do the analysis. The data will be first cleaned and prepared for analysis using various tool which we used in class. In addition, It will also involve formatting of date and time using Lubridate as it is a crucial part of the project. Some of interesting KPI's Calculated will be which is the busiest day of week for driver, which is the most bust month, which base has the highest number of trip, visualizing through heat map, the busiest boro in NYC. Another important part of project will be to get the weather data set and then join it with the Taxi data set and do the analysis, it will also help us answer the question how weather affects Taxi trips and help us know how it affects business for Drivers and also get some meaningful insights. This is my main idea but I might and more to as I get my hands on it. Both data sources will have cleaning, preparation, formatting and, editing.