Every time we google information, buy something online,make a new account, or simply listen to music online we are organically creating user data. As you can imagine there is a surplus of data in the world today. Data on stock prices, housing market, energy consumption, consumer market are just some areas with large amount of data available to the public. But how we use this data to make informed decisions is what is going to change the future. Those companies that have the ability to find trends and answer critical questions using their data are the ones who are going to prosper in the near future, those who can't will be left behind.

PHASE1

When thinking about how we were going to use the world of data to our advantage we originally planned on answering the question, "When is the best time to buy plane tickets?" However upon weeks of searching for the right data set and different projects to pivot towards we ran into two main problems; Kaggle data sets only had flight times and the Department of Transportation only had yearly fight prices. Therefore we decided to pivot and make the data set ourselves! Our objective was then to scrape flight data on a daily basis in order to create a workable database that we can share with others.

PHASE2

After pivoting we focused on learning how to scrape data from different sites. We first learned the basics from the web scraping lecture in class then focused on how we can apply those skills to our problem. We started basic scraping with Beautiful Soup and request and it worked because we were able to get all the HTML text with formatting. However we weren't able to get any relevant information because it was hidden behind Javascript. After that we attempted using Google Developers method and while it had an established library and easy to use commands, QPX express was not assigned to the project, furthermore there was a 50 query quota per day. Another downside to using QPX express is that it wouldn't be available after April 2018. We then again pivoted to using selenium and Phantom JS although it was difficult to set up on different computers it worked and we were able to scrape the data.

PHASE3

Once we were finally able to store the data, it was a matter of understanding how each query worked and then automating the task so that it could be saved into a csv file. First, users will provide a url to search for the flight they want. They specify the departure, destination, airline and date they want to fly within the url. Then, the function will pull all the information from that webpage and put it into a dataframe with

those factors along with the price of each flight. Furthermore, each time a new URL is entered into the function, a new dataframe will be automatically created with all the necessary information.

IMPROVEMENTS

If time allowed, there would be many more improvements to be made to our project. First, we would try to find the pattern in which Google parsed the flight data so that different dates of departure would work. Then, we would be able to come up with a huge dataset of many different destinations, airlines and price variations for many more time intervals. Another improvement that could be made was to automate this program to run on specific intervals, like every 15 minutes so that we could go back and answer our original question of what time of day is it cheapest to buy flights? Furthermore, we would've also made a user interface, for consumers like us who are data driven and would like to analyze how flights vary during times of day. What we have now is the basis of many popular cheap flight searching websites and if given more time and expertise, we could amount to uncovering trends in the flight industry which leads to transparency and a more equitable playing field for consumers.

TASK DIVISION

During the beginning of the project, we each contributed a fair share of time into market research to learn more about the flight industry. We researched factors that affect prices, typical pricing trends and other websites that have programs like the one we are trying to build. We all had different tasks with coding the program, but we would meet together bi-weekly to discuss outcomes, achievements, improvements and then add all the code to David's GitHub repository.

Deep Dave: Researching features of flight segments to collect which would impact ticket cost, as well as sources to collect the information from. First I worked on listing out flights and features to track/scrape. I worked on automation for the scrapping to repeatedly pull data for these flights from google flights.
David Lin: brainstorm planning for original project technical implementation and pivoting by discussing with Professor Sidhu, wrote code that bypassed Javascript loading and eventually made it into final project, implementing writing to CSV, organizing final codebase on Github
Sharon Ng: scraping via BS4 and QPX express, parsing data, table function and presentation slides
Vanessa Salas: research different scraping methods, analysis of data, Final Report & Presentation
Alexandre Vincent: