1. **(1 point) Explain the process of curating the data?**
   * Curating the data means that we clean and manipulate the data as needed, with the desired result being a more concise and complete dataset for further analysis. In the case of my dataset, I viewed the data in a Jupyter Notebook using a pandas data frame. My data was in 2 files with the same structure, so I had to concatenate the 2 files. After examining the number of unique values and any missing values in the data frame, I determined which columns provided useful insight and which ones were bulk that could be dropped. I also had to recreate some of the columns that were stored as hard to understand abbreviations so that they could be more easily understood by myself and any observers of my final project.
2. **(2 points) What tools and techniques are used in the process?**
   * I used the python packages pandas to create and manipulate my data in the form of a data frame. I used different pandas methods like dropna to drop missing columns, drop to drop unnecessary columns, unique to view the unique values in each column, concat to combine my 2 data sets, isnull and any to find columns with missing values, and map to create a new column based on given values in already existing columns. I also used the python package plotly, specifically the module plotly express, as this is my prefered plotting package. The plotly.express function "histogram" was used to view the distribution of my data.
3. **(2 points) What strategies were followed to cleanse the missing values and why?**
   * The only column that was missing data was called "shipping\_time". About 19.7% of the values in this column were null. I figured that this would be a very important independent variable in my later model, and I didn't want to feed any poorly imputed data into my model, so I dropped all of these rows that had missing values in this column. The data science phrase "garbage in, garbage out" was a guiding principle in this decision.
4. **(2 points) How many data attributes and records you ended up with after the cleansing process?**
   * After cleaning the data, I was left with 5114 records/rows and 9 columns/attributes.
5. **(1 point) Give the definitions of important data attributes.**
   * shipment\_id: This was a unique ID given to each shipment in the original data set. I debated whether or not to keep this in or simply refer to individual records with the data frame's index, but I ultimately decided that it may be useful to have this ID for further reference in the future.
   * freight\_cost: The cost/kg of the shipment.
   * gross\_weight: The weight in kg of the shipment.
   * shipment\_charges: This one was confusing to me at first, but looking back through the documentation on Kaggle, I learned that this is the minimum amount that a company would charge for a specific shipment. I may not end up using this in my final analysis, but I figured I would keep it in here just in case I do.
   * total\_cost: This was a field I calculated by multiplying the freight\_cost by the gross\_weight, as the freight\_cost is the charge per kg. This will be the dependent variable that I want to predict.
   * shipment\_mode: The way the product is shipped, either by air or ocean.
   * shipping\_company: One of 3 anonymous shipping companies used. The shipment\_charges field values are based on this field's value.
   * shipping\_time: Time in days that it took the shipment to reach it's destination.
   * sender: The origin country. This was originally stored in a field called source\_country with a 2-letter abbreviation as a value, so I calculated this field using the map function to make it easier to understand. However, all of the values here are "Great Britain," so I may end up dropping this field.
   * receiver: The destination country. This was originally stored in a field called destination\_country with a 2-letter abbreviation as a value, so I calculated this field using the map function to make it easier to understand.
6. **(2 points) Identify the dependent and independent variables related to the problem you are addressing?**
   * The independent variables will for sure be gross\_weight, shipment\_mode, shipping\_time, and receiver. Other possible independent variables may be shipment\_charges, shipping\_company, and freight\_cost, although these will likely be unnecessary.
   * The dependent variable that I am trying to predict is the total\_cost field.