1. **(2 points) What is exploratory data analysis? Why is it essential in a data science or data analytics project?**
   * Exploratory data analysis (EDA) is the process of investigating data to pull out some key insights, include which are the most important variables and whether or not any outliers exist within the data. EDA allows us to get a better understanding of relationships that may exist in our data before creating our models. It is essential in a data science project because it allows us to check our assumptions before getting to the more complicated model building practice, hopefully saving us some time and headaches at that point. If we begin creating and testing our model but don't understand our data, we may consistently get poor results or have to repeat our steps, whereas completing EDA should minimize and redundancies later in the process thanks to a better understanding of the data. Additionally, using EDA, we can spot any errors or outliers in the data and remove those before we get to the next steps. In my opinion, EDA allows us to perfect our data cleaning performed before, as we better understand our data and any flaws that may exist.
2. **(2 points) What are the various exploratory data analysis techniques? What specific techniques did you use in your project?**
   * Univariate EDA, either graphical or non-graphical (meaning with or without visuals, respectively) is where we examine one variable by itself. I performed univariate graphical analysis by creating histograms of each of my variables to better understand the distribution of each variable.
   * Multivariate EDA, either graphical or non-graphical, means that we examine multiple variables. I created a few scatter plots to better understand the relationship between some key variables. I also created a correlation matrix to visualize which variables had strong relationships to one another.
3. **(4 points) Explain the details of the techniques and the results of your exploratory analysis. (You may include a diagram to explain your exploratory analysis phase.)**
   * For my univariate graphical analysis, I examined each variable that had more than one value on it's own histogram. This allowed me to see whether or not data was skewed in any particular way. I found that most of the data was skewed to the left but with very long tails. I also created box plots and found that a lot of my variables had several outliers.
   * For my multivariate graphical analysis, I created a correlation matrix, which allowed me to see the relationship between variables. The general result was that there were two very strongly related variables, and the rest were not strongly correlated. I also created scatter plots, which allowed me to get a better understanding of how three variables related to one another thanks to an x-axis, a y-axis, and the ability to color points on the scatter plot. The correlation matrix can be seen below:
   * Using box plots (a form of univariate graphical analysis), I found that gross weight, total cost, and shipping time all had a lot of outliers, which was also apparent in the histograms. However, the histograms still looked somewhat normally distributed to me, so I figured that for the time being, it is okay to keep the outliers, as they may be relevant data points. However, when testing my regression model later, if the results are poor, I may have to come back and clean the data some more by removing these outliers.
   * I created scatter plots (a form of multivariate graphical analysis) and saw that of the two receiver countries, India received the highest cost and weight of shipments, the most expensive and heaviest shipments were shipped through ocean shipping as opposed to air, and anonymous shipping company one handled the heaviest and most expensive shipments.
4. **(2 point) What did you learn from this particular phase?**
   * Some interesting takeaways were that the gross weight of shipments in my dataset were somewhat normally distributed except for having a very long tail, most shipments cost $100,000 or less, air shipments were more common than ocean shipments in this data set, and most shipments took less than 6 days to reach their destination. Using multivariate graphical analysis, I found that the total cost of shipments was almost completely correlated to gross weight. Shipping cost had the next strongest correlation to total cost, but it was a week correlation.
   * Using box plots (a form of univariate graphical analysis), I found that gross weight, total cost, and shipping time all had a lot of outliers, which was also apparent in the histograms. However, the histograms still looked somewhat normally distributed to me, so I figured that for the time being, it is okay to keep the outliers, as they may be relevant data points. However, when testing my regression model later, if the results are poor, I may have to come back and clean the data some more by removing these outliers.
   * I created scatter plots (a form of multivariate graphical analysis) and saw that of the two receiver countries, India received the highest cost and weight of shipments, the most expensive and heaviest shipments were shipped through ocean shipping as opposed to air, and anonymous shipping company one handled the heaviest and most expensive shipments.
   * Overall, I learned that the most important independent variables will likely be gross weight, shipping time, shipping company, receiver and shipment mode. Total cost will remain the dependent variable. I will have to clean my data a bit more to drop some unnecessary columns.