Hi Kathleen,

After reviewing the data, I noticed several issues with the datasets:

Customer Demographics:

- Issues with consistency and formatting: the genders and states are written in multiple formats and creating duplicates when tables are merged with Customer Addresses. The formatting needs to be streamlined in order to properly process the data. There is also another unidentifiable value, 'U'. Please confirm if this stands for 'Unknown'.
- Issues with accuracy: I noticed that a value under date of birth was inputted correctly. It says that the oldest user was born in 1843.
- Issues with missing data:
    - There are users who are missing information regarding date of birth, tenure, job industry, and job titles. This information may be useful to have in determining and predicting target markets for purchasing bikes as we go into phase two of our analysis. These all point to potential salary ranges which could useful in determining what segments of customer buy bikes based on income.
    - Also, unlike the new customer list, there are no columns detailing rank or value. There is no explicit description describing what this is measuring and we will need Sprocket to clarify if this is a feature they would like to include. If so, we will need the rank and value descriptions as well as information for existing customers.
- Issues with relevancy: I noticed that there is a default column filled with unknown characters. Was this meant to be something else? If so, there may be an issue with how the raw data is being pulled in if we are missing relevant information.

Customer Addresses:

- Issues with consistency and formatting: The states are not formatted consistently. This formatting needs to be streamlined in order to properly process the data and avoid duplicates.
- Issues with missing data and relevancy:
    - Not all customers under the existing customer list have provided a postal code, or state which could have been useful in segmenting which locations are more likely to purchase bikes.
    - With that said, I don't think that it is relevant to have a customer's specific street number. We only really need a general location to get an idea of property values as this can potentially be an important feature in determining our market segment.

Transactions (3 months):

- Issues with consistency and formatting:
    - There are issues with consistency, particularly with multiple brands falling under the same product id or not being categorized at all. Knowing what brands are more generally in demand or maybe are popular by location would be an important feature to have in our analysis.

    - Similarly, missing data for online orders, brands, product line, and product size can be important features in determining general demand for a particular product or whether or not Sprocket should focus on online demand vs. in-store demand.
    - We have more than 3 month's worth of transactional data. Please specify which 3 month intervals our team is supposed to take a look at and analyze.
- Issue with formatting and relevancy: The product_first_sold_date column is not data formatted. I also don't believe that this column could be relevant as we care about the quantity and consistency of purchase. This column does not tell us if the customer continues to purchase or not beyond the first sold date.

New Customer List:

- Issues with consistency and formatting:
    - Property values are not in the same integer formats (ie. some values are written as 10.00 or 10 and they are being recognized as two individual data points). We will need to standardize the formatting on this. We also need to clarify what the unit of measurement is for the property value. Is this in the hundreds of thousands or millions.
    - As stated above under customer demographics, please elaborate and provide a description for what rank and value are measuring. This will help us determine it's relevancy.
    - Gender is also inconsistent and has different formats for Male/M and Female/F. We will need to standardize this as well.
    - Job industry is also missing data or is not applicable.

While there are quite a bit of issues on each table, we can reformat a lot of this information to be more consistent. We would just need Sprocket to elaborate and provide more information on some of the items I described above.

Thank you,

Justine