Dear customer,

I have looked carefully at the data provided (transactions, address and demographics for existing customers and demographic data + addresses for prospective customers). In general, it is adequately collected and presented with a few limitations detailed herein.

1. Transactions
   1. ~1,400 transactions for products with **product ID** of 0. It is unclear whether 0 is an actual product ID or is meant to indicate that the product is unknown. Action: *These entries are flagged for further exploration, but should not impact the downstream analysis.*
   2. There is a big gap between the last **customer ID** (at 5034) and the previous one at 3500. Action:*Again, this should not affect downstream analysis.* ***Recommendation***: *Ensure the customer IDs are unique and contiguous as adding new customers may cause some conflicts and/or loss of data.*
   3. There are missing values for:
      1. **Online order** (360)
      2. **Brand** (2910)
      3. **Product line** (197)
      4. **Product class** (197)
      5. **Product size** (197)
      6. **Standard cost** (197). Interestingly, there is also a gap in values between 1250 and 1500
      7. **Product first sold date** (197)

Action: *The 197 missing product-related entries are a very small proportion of the full data. Rather the corresponding entries, however, a model will be constructed to impute these values based on relationships with other available variables. It is of note that none of these variables are present in the new customer demographics data and are therefore not paramount for downstream analysis and model building*. Action: *A notable exception is* ***standard cost*** *as these values are required for the prediction of “profitable” customers. The values will therefore be imputed for this feature based on relationships and correlations with other features in the dataset.*

1. Customer Demographics
   1. There are 125 missing **First Name** entries. Action: *These will be replaced with ‘unknown’ and will not affect downstream analysis*. ***Recommendation:*** *It would be a good idea to obtain the first name of prospective customers as it will make reach-out more personable.*
   2. There are several designations for **gender** and a type for several values in the “Female” cateogry. Action: *All values were converted to read either ‘Female’ (‘F’, ‘Femal’ and ‘Female’), ‘Male’ (‘Male’ and ‘M’) or ‘U’ for unknown.*
   3. There are 87 missing values for **DOB** (date of birth). Action: T*his feature will be transformed to indicate years difference between each customer and the oldest customer in the dataset.*
   4. The oldest **DOB** was 1843-12-21. Action: *Looking at the transaction dates for this customer, compared with the transaction dates for other customers, I approximated the age of the customer to be around 1970. However, 1843 may simply have been a typo for 1943.*
   5. 2 entries where **deceased indicator** is “Y”. Action: *These entries should be eliminated from downstream analysis and modelling.*
   6. 87 missing values for **tenure**. Action: *As this would be an important factor in predicting high-value customers, the missing values will be imputed based on relationships with other features of the dataset.*
   7. 506 nulls for **job title** and 656 missing values for **job industry category** with only 105 values null for both. Action: *The missing values will be imputed to “unknown”.*
   8. There are essentially no values for **default.** Whatever values are not null are likely encrypted and illegible. Action: *The whole feature will be dropped out from downstream analysis.* ***Recommendation:*** *It would be great to make this information available as it could provide valuable insight into predicting high-value customers.*
2. Customer Address:
   1. All **state** designations were converted to abbreviations.
   2. **Address** does not correspond to **postcode**! Action: *The postcode was used to obtain coordinates as latitude and longitude, and the latter will be used for downstream analysis.*
3. New Customer Data:
   1. The **customer IDs** for new customers overlap with those for existing customers. Action: *Re-assign customer IDs following the customer IDs for existing customers (e.g. starting at 4004).*
   2. The **Rank** columns values are essentially duplicates of the customer ID and of the unnamed column preceding it. Action: *The unnamed column was removed.*
   3. 29 missing values for **Last Name**. Action: *Replaced with “unknown”.*
   4. 17 missing values for **DOB**. Action: *These values will be imputed based on other features of the dataset.*
   5. 106 and 165 missing values for **job title**  and **job industry category**, respectively. Action: *job title values will be replaced with “unknown” and job industry category null values will be imputed based on the other available features.*
   6. All values for **deceased indicator** are “N”, which brings little information to the dataset. Action: *The deceased indicator feature will be removed from downstream analysis,*
   7. **Address** does not correspond to **postcode**. Action: *Geolocation for each customer (latitude and longitude) will be done based on postcode. The rest of the features pertaining to customer location (address, state, country and postcode) will be removed.*
   8. There are **4 unnamed columns** (17, 18, 19 and 20 in the dataset). Their values are highly correlated (R2 ranging from 0.84 to 0.96). Action: *Since I do not have any information regarding what they pertain to, these features will be dropped after imputation of missing values for other features.*
   9. It is unclear how new customers were ranked (**Rank** feature).
   10. It is unclear what the feature **Value** refers to. Action: *It will be used for imputation of values missing from other features of the new customer data but not for downstream modelling and analysis.*
4. Merging all the existing customer tables based on customer ID, there are a total of 197 missing values for brand, product\_line, product\_class, product\_size, standard\_cost, date\_product\_first\_sold, 449 missing values for DOB, 2,397 for job\_title, 3,232 for job\_industry\_category, 3 for past\_3\_years\_bike\_related\_purchases, wealth\_segment, owns\_car and tenure, and 32 for address, postcode, state, country and property\_valueation.

Overall, there is enough data and reasonable ways to handle missing values in order to perform adequate analyses and build inferential and predictive models to determine high-value future customers.

Providing information about the 4 missing column labels in the new customer demographics sheet and an explanation of how these new customers were ranked may prove beneficial for data exploration, imputation and downstream analysis.

Best regards,

Vassil Dimitrov