Case Study Prep Notes

Primary Question: How do annual members and casual riders use Cicylistic bikes differently?

**Step 1: Ask**

Guiding questions

* What is the problem you are trying to solve?
  + The goal is to convert the ‘sometimes riders’ into regular accounts
* How can your insights drive business decisions?
  + If we can isolate key drivers through correlations or relatable market data, this can be used for further targeted research and marketing initiatives

Key tasks

* Identify the business task
  + The key business task is to identify the key and actionable differences between annual members and casual riders and present insights and hypotheses that can guide next steps
* Consider key stakeholders
  + Lily Moreno: Director of Marketing
  + Analytics Team
  + Executive Team
  + Customers: from perspective of data privacy and limitations on allowable data collection

Deliverable

* A clear statement of the business task
  + Identify key differences between annual membership accounts and casual rider purchases, then compile actionable insights and hypotheses that can guide next steps in the market expansion process

Step 2: Prepare

A screenshot of a computer

Description automatically generated

* Data will be stored in the Start\_Data folder, organized by year / quarter in file name, as CSVs.
* Final data is compiled as a data frame since it’s easier to do this with rbind once the data is cleaned (and formatting of the data values can be more easily managed this way)
* Licensing is not an issue in this case since the data is coming from the class resources, user data is not included in the dataset and therefore not a concern here
* Data integrity was investigated using dplyr::count(df, x) methodology (and some charting for confirmation)
  + birthyear field has 791,216 NA values
  + gender field has 791,280 NA values
* Out of 2,454,634 records, this is min

Process

A screenshot of a computer screen

Description automatically generated

* Data integrity checked, decision taken to not use gender or birth year for meaningful conclusions (missing data on 32.2% of records)
* Will be doing the analysis in R for practice, could also be excel but calculation times would just be painful
* Cleaning process documented
* Added ride\_length and day\_of\_week fields to the data frame

A screenshot of a computer screen

Description automatically generated

A screenshot of a questionnaire

Description automatically generated

* Did a count of ride length, found top 10 values (pic below)

A screenshot of a computer

Description automatically generated