McGill | School of Continuing Studies | École d'éducation permanente

Assignment: Using Alteryx to Explore BIXI Data - Data Science for Business Decisions

Student ID: 261032796
Name: Ahmed Ibrahim
Email: ahmed.ibrahim6@mail.mcgill.ca

SET 1: Descriptive Answers

1. **$[R_T = 4926118.05\ CAD]$**

Descriptions

To estimate BIXI's 2019 revenue, two datasets were utilized. The first dataset consists of the year 2019 trip history (*biximontrealrentals2019* folder). The second one consists of a chart: "*Number of purchases of Memberships and Short-term Access*." Both can be found in the Open Data section of BIXI's website. The chart data has been converted to table 1, but only the membership purchases remain relevant for the herein analysis.

Table 1. *The number of membership and occasional purchases from BIXI clients in 2019. The data was extracted from a bar chart from the Open Data section of https://bixi.com/en/page-27.*

Month	Membership Purchases
January	21
February	36
March	4541
April	13488
May	10620
June	6697
July	5622
August	3817
September	2709
October	920

The pricing system used for cyclists depends firstly on whether they are BIXI members or occasional cyclists. Members can pay a flat fee for seasonal usages covering the entire period of bike riding for 2019. Each type of cyclist pays additional fees after a specific duration. They pay a rate of 5 cents per minute after each minute pass the 45-minute mark for members. The occasional cyclists pay a higher rate after 30 minutes at a rate of 10 cents per minute for those without an OPUS card, but the OPUS cyclists pay 9 cents per minute. Thus, each trip can come from 4 different customer groups such that the seasonal

members, the monthly members, one-way cyclists, and the OPUS one-way cyclists (see Assumption 1 to 15). The pricing system is summarized in table 2.

Table 2. The pricing system for the BIXI 2019 revenue analysis. The overage rate for members applies to rides longer than 30 minutes and 45 minutes for one-way cyclists. The data for this table is found at https://www.bixi.com/en/pricing.

Type of Member	Seasonal Fee (CAD)	Monthly Fee (CAD)	Flat Fee (CAD)	Overage Rate (CAD/min.)
Seasonal Member	99.00	0.00	0.00	0.05
Monthly Member	0.00	19.00	0.00	0.05
One-Way Pass	0.00	0.00	0.50	0.10
OPUS One-way Pass	0.00	0.00	0.45	0.09

In this analysis, there are two sources of revenue which are the revenue generated from memberships and occasional trips, as seen in the following equation:

$$R_T = R_M + R_O \ [1]$$

R_T is the total generated revenue, R_M is the membership revenue, and R_O is the revenue from the occasional trips. The revenue from memberships can be represented by the following:

$$R_M = R_{SM} + R_{MM} \ [2]$$

R_{SM} is the revenue from seasonal memberships, and R_{MM} is the revenue from monthly memberships. Both revenues can be computed with the following:

$$R_{SM} = 99N_{SM} + R_{SMOF} \ [3]$$

$$R_{MM} = 19N_{MM} + R_{MMOF} \ [4]$$

N_{SM} is the total number of seasonal memberships purchased, and R_{SMOF} is the revenue generated by the overage portion of the seasonal member trips. For equation 4, N_{MM} is the total number of monthly payments, and R_{MMOF} is the revenue generated by the monthly member trips' overage portion. The overage portion of both equation 3 and 4 can be found below:

$$R_{SMOF} = 0.05T_{SMOF} \ [5]$$

$$R_{MMOF} = 0.05T_{MMOF} \ [6]$$

T_{SMOF} is the sum of minutes over 45 for each trip from seasonal member trips, and T_{MMOF} *is* the equivalent for the monthly member trips. Since the overage rates for both member types are the same, equations 5 and 6 can be combined as follows:

$$R_{MOF} = R_{SMOF} + R_{MMOF}$$

$$R_{MOF} = (0.05T_{SMOF}) + (0.05T_{MMOF})$$

$$R_{MOF} = 0.05[T_{SMOF} + T_{MMOF}]$$

$$R_{MOF} = 0.05T_{MOF} \; [7]$$

R_{MOF} is the revenue generated by the overage portion of all member trips, and T_{MOF} *is* the sum of minutes over 45 for each trip from all member trips. By combining equation 3, 4, 7; an alternative form of equation 2 can be obtained with the following:

$$R_M = R_{SM} + R_{MM}$$

$$R_M = (99N_{SM} + R_{SMOF}) + (19N_{MM} + R_{MMOF})$$

$$R_M = 99N_{SM} + 19N_{MM} + R_{SMOF} + R_{MMOF}$$

$$R_M = 99N_{SM} + 19N_{MM} + R_{MOF}$$

$$R_M = 99N_{SM} + 19N_{MM} + 0.05T_{MOF} \; [8]$$

The revenue from the occasional trips can be calculated with the following:

$$R_O = R_{OWT} + R_{OOWT} \; [9]$$

R_{OWP} is the revenue from the one-way pass trips without discounting, and R_{OOWP} is the OPUS one-way trips. Both can be calculated with the following:

$$R_{OWT} = 0.50N_{OWT} + R_{OWTOF} \; [10]$$

$$R_{OOWT} = 0.45N_{OOWT} + R_{OOWTOF} \; [11]$$

N_{OWT} is the total number of one-way trips without discounting, and R_{OWTOF} is the revenue generated by the overage portion of these one-way trips. For equation 9, N_{OOWT} is the total number of OPUS one-way trips, and R_{OOWTOF} is the revenue generated by the overage portion of those OPUS trips. The overage portion of both equation 3 and 4 can be found below:

$$R_{OWTOF} = 0.10T_{OWTOF} \; [12]$$

$$R_{OOWTOF} = 0.09T_{OOWTOF} \; [13]$$

T_{OWTOF} is the sum of minutes over 30 for each trip from the one-way trips, and T_{OOWTOF} is the equivalent for the OPUS one-way trips.

Before starting the revenue analysis, the trip data needed some minor prepping (see step 1 of calculations). The first source of revenue computed was the one generated from the overage fees of member trips. Since the duration of member trips was known (see Assumption 15), it was possible to compute R_{MOF} (see step 2), and the value obtained was 14953.90 CAD. Two topics needed to be addressed to solve the rest of the equations: the distribution of monthly vs. seasonal membership purchases and what portion of occasional trips are from OPUS card users?

To address the first question, membership purchases from January to May were labeled as seasonal and for the rest of the months as monthly (see assumptions 16 and 17). Half of the occasional trips were assumed to be from OPUS cardholders (see assumption 18).

The total number of monthly payments made was found to be 73762 and 28706 for the seasonal payments. As a result, the total revenue generated from members was computed to be 42583325.90 CAD.

The total number of occasional trips was tabulated to be 1005062, and the total amount of overage minutes was found to be 2004081 minutes. Knowing that half the fares are discounted since half of the trips are assumed to be from OPUS users, the revenue generated from occasional trips was 667792.15 CAD.

The revenue generated from BIXI 2019 using the herein assumptions is 4926118.05 CAD.

Assumptions

The estimate for 2019 uses the following assumptions:

1. The estimate is for gross sales, which means there is no consideration for any sources' expenditures.
2. The herein estimate does not include sources of revenue beyond their bike-share service.
3. The only biker packages accounted for in the analysis are seasonal memberships, monthly memberships, the one-way pass, and the OPUS one-way pass.
4. In 2019, BIXI hosted its 4th annual "Manulife's Free BIXI Sundays." Thus, for May 26, June 23, July 28, August 25, September 29, and October 27, the generated revenue from non-members is assumed to be null for the enumerated dates.
5. BIXI offered numerous promotions to acquire and retain customers. Thus, the group memberships, corporate promotions, student promotions, and other related promotions are omitted from the analysis.
6. The BIXI Amis point system program did not account for in the analysis.
7. BIXI offered various promotions before 2019 that may have been carried over to loyal customers; these scenarios are omitted from this analysis.
8. The pricing system used in the analysis is based on the current edition (February 2021).
9. It is assumed that BIXI generated no bike ride revenue for January, February, March, November, and December of 2019.
10. All rides are assumed to be on non-electric BIXI bikes.

11. All memberships are assumed to start on the first day of the month.
12. It is assumed that they are not discounts for memberships in this analysis.
13. The data from table 1 is assumed to be voided of missing values.
14. The trip dataset is assumed to be voided of duplicates and missing values.
15. When converting the duration of trips from seconds to minutes, the minutes are rounded to the nearest unit. Thus, it is assumed that the approximation will have a negligible impact on the final estimation.
16. As for the membership revenue, the customer churn rate is assumed to be null. Thus, customers that sign up for a monthly membership stay committed for the remainder of the season. Assuming members pay effectively, memberships purchase from June to October will be labeled as monthly memberships. For instance: if someone purchased a membership in May 2019 and wanted to remain a customer for the remainder of the biking season, a seasonal membership would still cost 99$, but the sum of all monthly payments would be 114$. Thus, only the membership purchases from January to May are considered seasonal purchases.
17. In table 1, there is no indication that if a user is a monthly membership that their monthly payments are considered one purchase or each monthly payment as singular payments. In this analysis, a monthly membership purchase by a cyclist will represent their monthly payments' culmination.
18. OPUS card users can use the card to board the bus or train services in Montreal. Based on the BIXI website, only cyclists with valid OPUS cards are eligible for the OPUS discount. It will be assumed that half of the occasional cyclists were from OPUS card users.

Calculations

Step 1: Data preparation for trip data.

The trip data folder contains 8 comma-separated values (CSV) files, where 7 are the tabulated trip data for each month, and the last file contains geospatial data for each BIXI station. For the sake of the revenue analysis and the cumulation of assumptions, the station CSV file (*Stations_2019*) will not be necessary. For each monthly dataset, each recorded trip's fields include the start dates, start station codes, end dates, end stations codes, duration of trips in seconds, and a Boolean to label trip as a member or occasional trip. For this analysis, only the duration of trips and the member Boolean were retained. All the 7 CSV files for the monthly trips have been inputted in Alteryx and unionized with the "union" tool. Using the "record" tool, the trips have been indexed. Using the "select" tool, only the following data types have been checked: "RecordID," "duration_sec," and "is_member." These fields' data types have been changed respectively from strings to Int64, Int64, and Byte. An "output" tool was used to generate a CSV file

containing the prepped trip data. The Alteryx file name for this step is *TripDataPreparationWorkflow.yxmd,* and the generated CSV file is called *2019combinedBIXItripdata.csv*.

Step 2: Calculate the revenue generated from the overage fees of member trips

A new Alteryx workflow was created for this step called: *RevenueFromMembersOverageFees.yxmd*. The three fields' data types are set to be the same as the ones used in step 1. Since the overage rates are set in Canadian dollars per minute, the durations in seconds must be converted to minutes. Using the "formula" tool, a new column called "duration_min" was generated containing the duration of trips in minutes. The total number of trips stands at 5597845 trips, whereby using the "filter" tool (is_member = 1), 4592783 trips are from members. To find the trips longer than 45 minutes, another "filter" tool was used (duration_min > 45). Only 23526 trips remained. Since the overage charges start to pass the 45-minute mark, a new column can be generated containing the difference between the "duration_min" and the threshold of 45 minutes. The new column is called "overage_duration_min." Using the "summarize" tool and its "sum" action, the sum of minutes over 45 for each trip from all member trips (T_{MOF}) was found to be 299078 minutes. Using equation 7, the revenue generated by the overage portion of all member trips can be found below, and the same formula was used in Alteryx with the "formula" tool:

$$R_{MOF} = 0.05T_{MOF}$$

$$R_{MOF} = 0.05(299078\ min)$$

$$R_{MOF} = \ 14953.90\ CAD$$

Step 3: Calculate the revenue generated from the members

Using equation 7, two values need to be calculated to compute the revenue generated from the members. The data from table 1 and the assumptions associated with this revenue are used. A new workflow has been created for this step called: *RevenueFromMemberswithoutOverageFees.yxmd*. Table 1 data has been inputted in that Alteryx workflow, and the rows have been indexed like step 1. Using the "select" tool, the membership purchases field's data type has been set to Int64, and the occasional purchases field was removed from the output dataset. The record IDs can represent the numerical values for months to separate the rows for the seasonal and monthly memberships in this table. The "filter" tool was used, where if a record ID is smaller than 6, it goes in the seasonal membership table. Otherwise, it was filtered to the monthly membership table.

For the monthly branch, the number of months remaining for a purchase needed to be calculated. For monthly members who bought in June, they would have 5 months remaining in the memberships, including the month of June, where they paid each month's monthly rate. Using the "formula" tool, a column can be added called "months_remaining," where the remaining months can be calculated by subtracting 11 by the record ID of a given row. The total number of monthly payments for each start month can be calculated by finding the product of the monthly membership purchases and the months remaining. This was done using the "formula" tool, where a new column was generated titled

"total_number_of_monthly_payments_for_a_given_month". Using the "summarize" tool similarly in step 2, the total monthly payments (NMM) were computed to be 73762.

For the seasonal branch, the sum of seasonal purchases can be found using the "summarize tool" with the "sum" action. All the seasonal purchases (N_{SM}) from January to May are added up, and the result is 28706.

By using equation 8 and knowing R_{MOF} is equal to 14953.90 CAD, N_{MM} is 73762, and N_{SM} is 28706; the revenue generated (R_M) by members can be computed:

$$R_M = 99N_{SM} + 19N_{MM} + 0.05T_{MOF}$$

$$R_M = 99(28706) + 19(73762) + 14953.90$$

$$R_M = 4258325.90\ CAD$$

Step 4: Calculate the revenue generated from the occasional trips

Based on the assumptions, half of the occasional trips are known to be from OPUS users. Thus, the number of occasional trips for each is equal, and the total of overage minutes is assumed to be equal as well. A new workflow has been created called: *RevenueFromOccasionalTrips.yxmd*. To start the workflow, the workflow from step 2 has been repurposed for this step. Exactly like step 2, only three fields are necessary: the record IDs, the duration in seconds, and the Boolean to check if the trip is from a member. However, at the first "filter" macro, the branch at the true output results from the rows with a null value for its "is_member" boolean represents the occasional trips. As a result, the number of occasional trips was determined to be 1005062. Thus, N_{OWT} and N_{OOWT} were both equal to 502531 trips. Another "filter" tool was used to find the trips with overage time, where only the rows with durations longer than 30 minutes were kept in the true output. Out of 1005062 occasional trips, 126969 were longer than 30 minutes. Each row's overage duration was calculated using the "formula" tool by subtracting 30 minutes from each duration in minutes. Using the "sum" action from the "summarize" tool, the total amount of overage minutes was found to be 2004081 minutes. Thus, T_{OWTOF} and T_{OOWTOF} were equal to 1002040.50 minutes. Using equations 9 to 13, it is possible to compute the revenue from the occasional trips:

$$R_O = R_{OWT} + R_{OOWT}$$

$$R_O = (0.50N_{OWT} + R_{OWTOF}) + (0.45N_{OOWT} + R_{OOWTOF})$$

$$R_O = 0.50N_{OWT} + (0.10T_{OWTOF}) + 0.45N_{OOWT} + (0.09T_{OOWTOF})$$

$$R_O = 0.50(1005062) + 0.10(1002040.50) + 0.45(1005062) + 0.09(1002040.50)$$

$$R_O = 667792.15\ CAD$$

Step 5: Calculate the total generated revenue

The total revenue generated can be found by using equation 1 and the values obtained from step 2 and 4:

$$R_T = R_M + R_O$$

$$R_T = (4258325.90\ CAD) + (667792.15\ CAD)$$

$$R_T = 4926118.05\ CAD$$

2.

a. .

The 2019 BIXI dataset consists of trip datasets for each month that bike trips were recorded. The fields include start dates, start station codes, end dates, end station codes, duration of trips in seconds, and a Boolean to label trip as a member or occasional trip. Complimentary to these datasets, there is also a table containing more details on the stations. The station codes from the trip datasets are used as key values to match specific rows that provide three fields for each station, such as the station's name, latitude, and longitude. If BIXI intends to build a model that predicts each trip, it is assumed that the trip's starting point and the endpoint will be at BIXI stations.

Since the trip duration is the target variable, it would be worthwhile for the BIXI analysts to access training and validation data. Since BIXI is already tracking each trip's duration in seconds, this field can be used as the basis for building the desired forecast model. Additionally, each trip's start date and end date are tracked, including the start time and end time to the nearest second for the same trips. These two fields are crucial as the trip's duration can be calculated by the absolute difference between the start time and end time. As a result, BIXI can either use these two timestamps to calculate the duration of the trip or opt to measure the duration of the trip in real-time with some electronic apparatus. By recording each trip's year, the duration of trips can be compared to recorded durations of previous years, and patterns can be deciphered. As for the months being tracked, this can be used to discover unique features that are month-dependent and relate to the duration of trips (ex: weather conditions, precipitations). The days being recorded would allow BIXI to follow daily trends (ex: usage during holidays, weekday vs. weekends). Recording the time would enable them to detect fluctuation of trip times during the day, which can be due to traffic on the roads, the lack of traffic (ex: riding bikes during the night), or other features.

The trip data contains records of whether a member or non-member made a trip. This will enable BIXI to compare the duration of trips from members vs. occasional cyclists. Features unique to each group can be identified to predict the duration of trips better. Naturally, the duration of trips may have higher standard deviations for the occasional cyclist subpopulation than the ones from the member subpopulation. This may even lead to having different models for forecasting the duration of the trip for an occasional ride or member ride.

The start and end station codes for each trip would allow the analysts to match a specific trip duration to a particular route. Since the choice of route, starting point, or destination will have without a doubt an influence on the duration of a trip and it must be tracked for each trip. However, as for the table with the station info, this may not be useful for predicting the duration of trips since the station codes can be used to differentiate the stations within the trip datasets. On the other hand, the coordinates of these stations

(latitude and longitude) can be used to fetch geospatial data from external sources, which can be used to find additional features for build the herein predictive models.

In conclusion, all the fields in the trip dataset are useful, but the station info may not be as beneficial without using other datasets to forecast the duration of trips.

b.

The route's environment can influence a trip's duration, the bike's condition, and the cyclist.

The environment of the route can have many different variables that can influence the trip of a cyclist. For starters, the weather is a feature to account for when cycling. A rainy day renders smoother surfaces on the roads, which can make the bike trips longer. A hot and dry summer day may lead the cyclist to take breaks more frequently for hydration, making the trips longer. When it comes to cycling, there is a famous saying by USA Cycling that goes as follows: "there is no such thing as bad weather, only bad gear." Thus, a cyclist who is mindful of the weather conditions can minimize the effects of poor weather during their ride, but weather features remain worthwhile for the sake of building this model. With a Weather API and the station info, the precipitation and temperatures can be tracked for each trip.
Furthermore, forecasted weather data can be fed to the model to adjust the model's predicted trip outputted duration. Another relevant environmental condition consists of the possibility of obstruction in the routes. These road hazards can be due to on-going construction or accidents. Using the stations' coordinates, the potential routes can be generated using the Google Maps API, which in turn hazard data can be fetched to train the model to adjust the predicted duration of trips depending on the route conditions. Similarly, another factor that may affect the target variable is the traffic condition on the routes. To account for this, the Google Maps API or other geospatial APIs can be used to fetch data on the traffic conditions for the top routes connecting the starting point to the destination.

As for the bike's condition, a bike being well maintained or having some mechanical issues will affect the cyclist's riding experience and the duration of their trip. As a result, it would be worthwhile tracking what bike was used on that trip. Since all bikes have a bike ID, these IDs can be tracked in the trip data. Beyond being used for predicting the length of a trip, tracking the bike ID in correspondence to each trip may enable early detection of defective bikes.

All cyclists are not built the same. Thus, the physical conditioning, state of mind, height, weight, gender, and other physiological features of the cyclist can influence their trip duration. These demographics can be obtained during the cyclist's sign-up process or by fetching data from applications on their cellular devices upon their permission. With some data explorations, the useful cyclist features can be identified to be tracked for predicting the duration of a given trip.

In addition to the currently tracked fields by BIXI in their trip datasets, it would be worthwhile for BIXI to include features related to the road and weather conditions of the routes, physical traits related to the

cyclists, and the ID of the bike used for that given trip. These mentioned features would be interesting to investigate to create a robust model to predict the duration of trips from BIXI's various stations.