Google Data Analytics Capstone

Case Study 1: CYCLISTIC BIKE-SHARE

TOOLS USED: Excel, Microsoft SQL Server, Tableau, RStudio

Scenario

You are a junior data analyst working in the marketing analyst team at Cyclistic, a bike-share company in Chicago. The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore, your team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, your team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations.

Ask

At this phase, I asked myself, "How do yearly members and casual riders use Cyclistic bikes differently?". to help me with my analysis. The head of marketing, Lily Moreno, thinks yearly Members are significantly more profitable than casual riders. As a junior data analyst at Cyclistic, I am responsible for assisting the marketing team in analyzing the distinctions between Casual riders and Members to provide recommendations that will optimize profit by converting Casual riders into yearly Members.

Prepare

I saved the dataset that was downloaded for this case study. I had to download information from the previous 12 months. I downloaded travel information from April 2020 to November 2022. It is a public dataset that Motivate International Inc. has made available.

Dataset: Index of bucket "divvy-tripdata"

Process

The various files were integrated into a single Excel worksheet. The data was then cleaned by eliminating duplicates and rows with ride start times later than ride finish times (i.e., started at > ended at). I loaded the data into a Microsoft SQL server to speed up the data-cleaning process. I eliminated rows where columns like station id and station name contained NULLS. Using Power Query to combine the 12 tables into one, I exported the data back into Excel.

Analyze

In Excel, I made the ride length and day of week columns to determine the length of the journey and record the day of the week the ride was taken. I used pivot tables to aid in my data synthesis. I then switched to RStudio to build data frames and use ggplot2 to display the data. I used R's mean, median, max, and min functions to compute and summarize the data.

Share (Data Viz)

Tableau: https://public.tableau.com/views/Cyclisticbike

share_16721336575550/Cyclisticbike_share?:language=en-

US&:display_count=n&:origin=viz_share_link

R, Excel: SolarStellaris/Google-Data-Analytics-Case-Studies (github.com)

Act

In this final step, my task was to state and act on key findings by providing recommendations that will help the marketing director (my manager) maximize profits.

Key Findings

1. Casual riders were found to take longer trips or rides on average than annual Members. This could mean Casual riders used Cyclistic bikes for leisure.

2. Casual riders often ride on weekends whereas annual Members use the program more over the week than on weekends. This could indicate that annual Members are using the bikes to commute to work.

3. Casual riders preferred using electric bikes to other ride types like classic and docked bikes. Annual Members used classic bikes more than other types. The reason Casual riders mostly rode electric bikes could be because they rode longer so, electric bikes eased propulsion.

Deliverables (Recommendations)

1. Users that take long rides could be offered a discount on renting bikes or better yet annual Members that travel more than a certain distance say for instance 10 miles should be offered a discount.

2. Cyclistic bike-share should introduce a weekend only membership that costs less than the current 7-day membership.

3. A rebate program in form of refunds to annual Members that use a certain ride type more others should be introduced. It could be a small percentage of the cost.

Case Study 2: BELLABEAT

TOOLS USED: Excel, Microsoft SQL Server, Tableau, RStudio

Scenario

Bellabeat, a high-tech company that manufactures health-focused smart products. Bellabeat product were beautifully designed and inspiring. Bellabeat collecting data on activity, sleep, stress, and reproductive health to empower women with knowledge about their own health and habits.

Ask

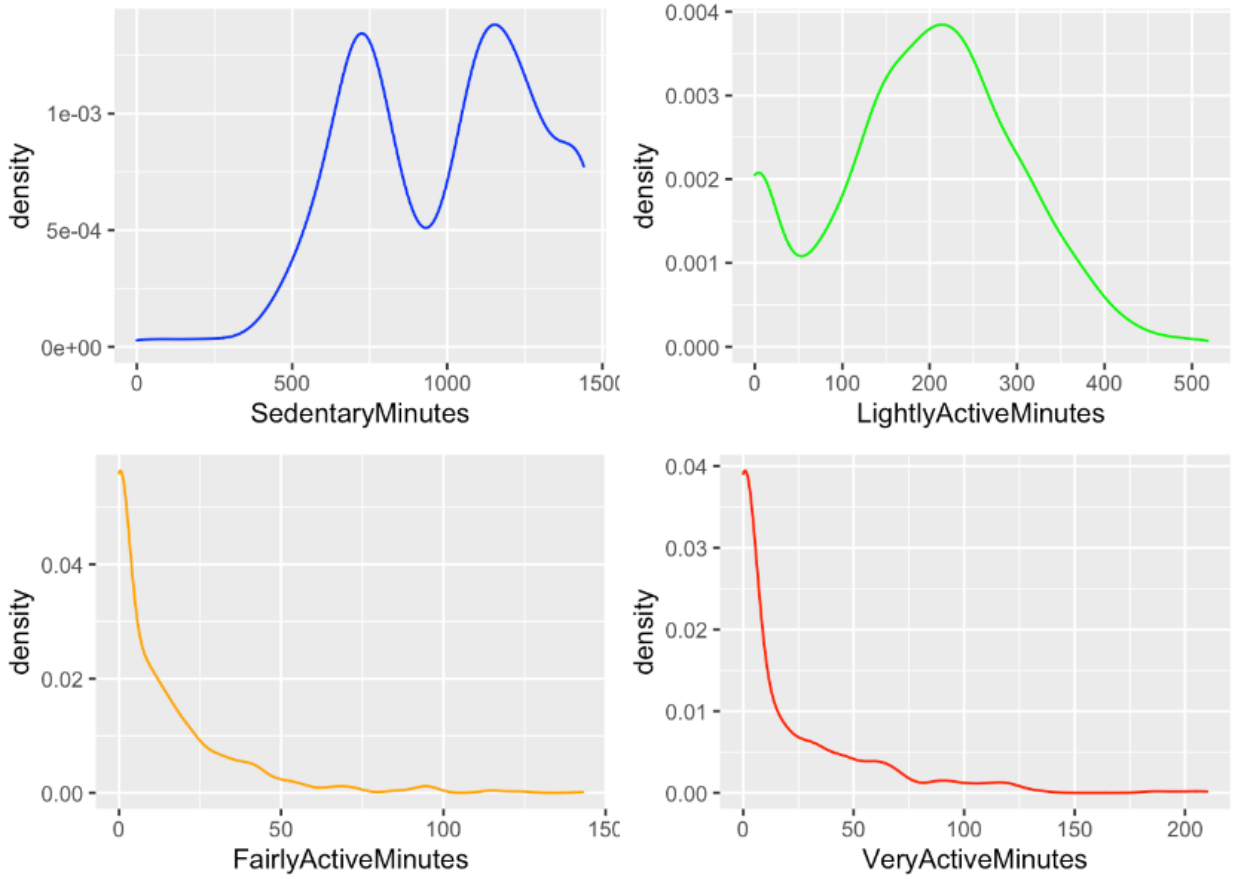
This pilot report will provide the answers to the following queries. What features do non-Bellabeat smart devices offer? How do non-Bellabeat smart devices perform on similar features? What use patterns are there for these features? How do consumers rate them? Apply the findings from the data collected from Bellabeat devices.

Prepare

Bellabeat offers the following products: the Bellabeat app, Leaf, Time, Spring, and Bellabeat Membership (the Membership). Through Mobius, we obtained the FitBit fitness tracker data. We also offer Fitabase Data Dictionary, which contains the information for this open dataset. There are 18 datasets, each including different participant activity and fitness information records. Load some datasets, then gradually become comfortable with them.

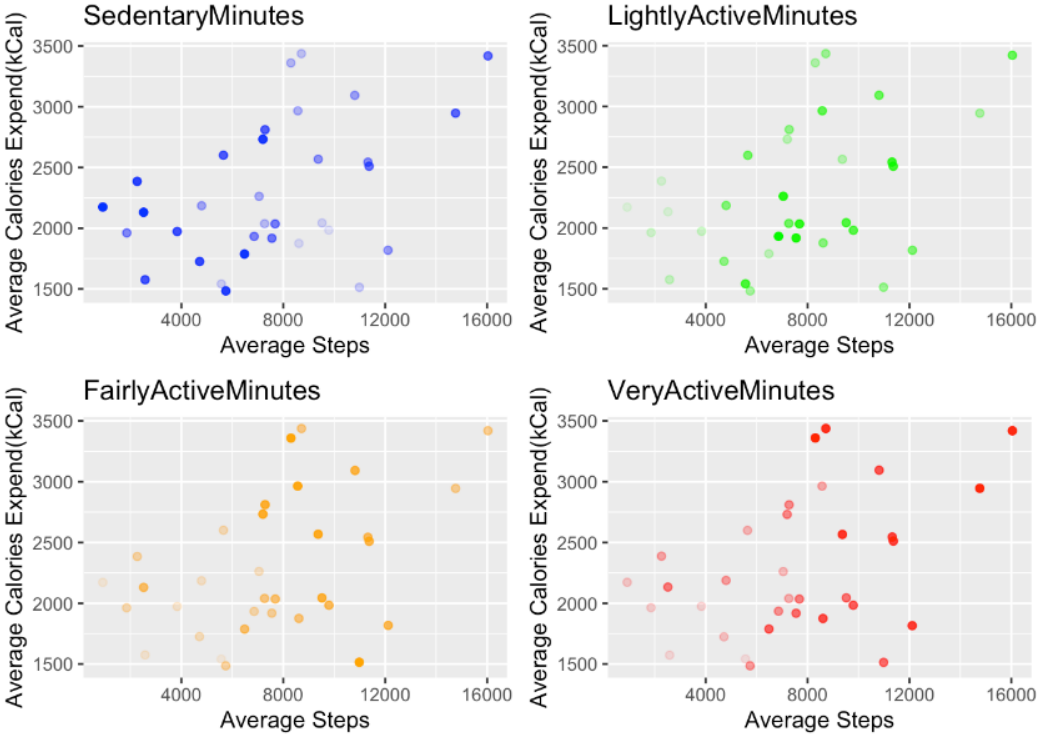
Process and Analyze

Based on these descriptive stats, we see the distribution of sample stats are most likely not normal. We further plot a density plot to demonstrate the distribution of activity minutes.



The activity minute records are not normal. However, to simplify this analysis, we assume a normality based on *Central Limit Theorem*. Graph a correlation plot with daily average activity minutes for each individual.

	TotalSteps	VeryActiveMinutes	FairlyActiveMinutes	LightlyActiveMinutes	SedentaryMinutes	Calories
TotalSteps	1.00	0.67	0.50	0.57	-0.33	0.59
VeryActiveMinutes	0.67	1.00	0.31	0.05	-0.16	0.62
FairlyActiveMinutes	0.50	0.31	1.00	0.15	-0.24	0.30
LightlyActiveMinutes	0.57	0.05	0.15	1.00	-0.44	0.29
SedentaryMinutes	-0.33	-0.16	-0.24	-0.44	1.00	-0.11
Calories	0.59	0.62	0.30	0.29	-0.11	1.00



Looking at the scatter plot, color opacity stands shows the magnitude of average activity minutes. Higher opacity indicates more activity minutes. Average daily steps and average daily calories expenditures are likely to have a linear relationship. Moreover, correlation plot shows a weak negative correlation for sedentary minutes with both daily steps and calories expenditure, and medium correlation for 3 other active minutes with both daily steps and calories expenditure. We could conduct similar analysis on other three datasets.

Share (Data Viz)

R: SolarStellaris/Google-Data-Analytics-Case-Studies (github.com)

Key findings

1. The fitbit data have some serious issues. It dates back to 2016, has many missing data, and a rather small sample size.
2. Based on limited data, participants were most likely to complete activity tracker data. Sleep data and heart rate data were not that easy to keep track everyday. Weight log is the least complete data.
3. Activity pattern. Daily average active minutes were 21.16(very active), 13.56(fairly active), 192.8(lightly active), 991.2(sedentive). Total steps and total calories were positive correlated. Daily active minutes are positive correlated with total calories.

Act

Final Recommendation to Bellabeat:

1. When collect Bellabeat users' data, make sure we have enough sample size, longer periods and high quality data.
2. Metadata is important, and specially critical for distinguish tracker record data and user record data.
3. We only analyzed tracker data so far. Bellabeat should incorporate all assets to draw a multidimisional user profile.
4. Bellabeat should consider using their membership as motivation to encourage user participate in activities and records.