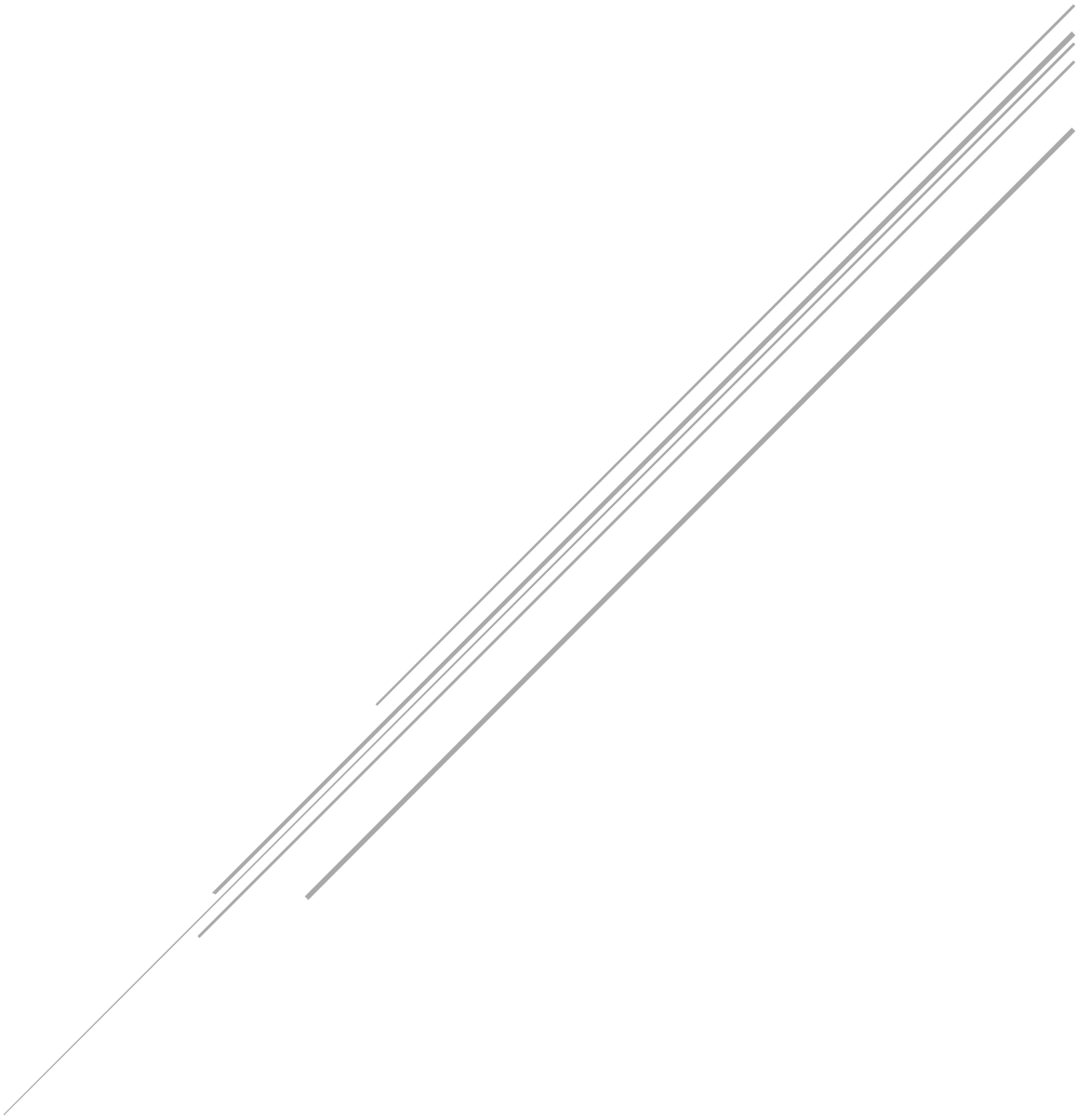
DATA ANALYST PORTFOLIO

Case Studies & Projects



Waleed El-Damaty

Table of Contents

How Does a Bike-share navigate speedy success?

Executive Summary

Our case study looks at Cyclistic, a successful bike-share company in Chicago. We're going to analyze historical bike trip data to identify trends. The outcomes of the analysis shall aid in understanding better how annual members and casual riders differ. This will help to design proper marketing strategies aimed at converting casual riders into annual members thus maximizing the number of annual memberships securing the company's future success.

Phase 1- Ask

At this stage of the data analysis, key stakeholders are identified and questions are posed to guide the process.

Introduction

In 2016, Cyclistic launched a successful bike-share offering. Since then, the program has grown to a fleet of 5,824 bicycles that are geo-tracked and locked into a network of 692 stations across Chicago. Until now, Cyclistic's marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the flexibility of its pricing plans: single-ride passes, full-day passes, and annual memberships. Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members. Director of Marketing Lily Moreno believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, Moreno believes there is a very good chance to convert casual riders into members. Thus, the main goal is to design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the marketing analyst team needs to better understand how annual members and casual riders differ.

Products

1. Cyclistic: A bike-share program that features more than 5,800 bicycles and 600 docking stations. Cyclistic sets itself apart by also offering reclining bikes, hand tricycles, and cargo bikes, making bike-share more inclusive to people with disabilities and riders who can't use a standard two-wheeled bike. The majority of riders opt for traditional bikes; about 8% of riders use the assistive options. Cyclistic users are more likely to ride for leisure, but about 30% use them to commute to work each day.

The Stakeholders

1. **Lily Moreno**: The director of marketing and your manager. Moreno is responsible for the development of campaigns and initiatives to promote the bike-share program. These may include email, social media, and other channels.
2. **Cyclistic marketing analytics team**: A team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy. You joined this team six months ago and have been busy learning about Cyclistic's mission and business goals — as well as how you, as a junior data analyst, can help Cyclistic achieve them.
3. **Cyclistic executive team**: The notoriously detail-oriented executive team will decide whether to approve the recommended marketing program.

Business Task

In this case study, we will analyze the historical bike trip data to identify trends. The outcomes of the analysis shall aid in understanding better how annual members and casual riders differ. This will help to design proper marketing strategies aimed at converting casual riders into annual members.

Business Questions to Answer

1. How do annual members and casual riders use Cyclistic bikes differently?

Phase 2 - Prepare

In the Prepare phase, we identify the data being used and its limitations.

Data Source

The Data set is publicly available through this link (Download the previous 12 months of Cyclistic trip data here). The data has been made available by Motivate International Inc. under this license. The data contains 20 data sets that show the riders' patterns from April 2020 to November 2021.

Data Limitations

Due to the fact that this was internal data, we can assume that it is authentic and impartial, although I did notice that some rows in the (ended_at) column occurred before the (started_at) column. The (ride_id) was used instead of the (rider_name) to protect the riders' privacy. However, showing the total amount of money spent per ride could've helped in identifying the variation between annual and casual members in terms of spending. Additionally, some of the stations' names were missing as well.

Data Quality

A good data set is usually reliable, original, comprehensive, current and cited. However, in our case, the data set is considered to be of a bad quality due to the following reasons:

1. It is reliable: Collected internally
2. It is original: Collected internally
3. It is comprehensive: Cover 20 months' worth of data
4. It is current: Updated and relevant
5. It is properly cited: Collected internally

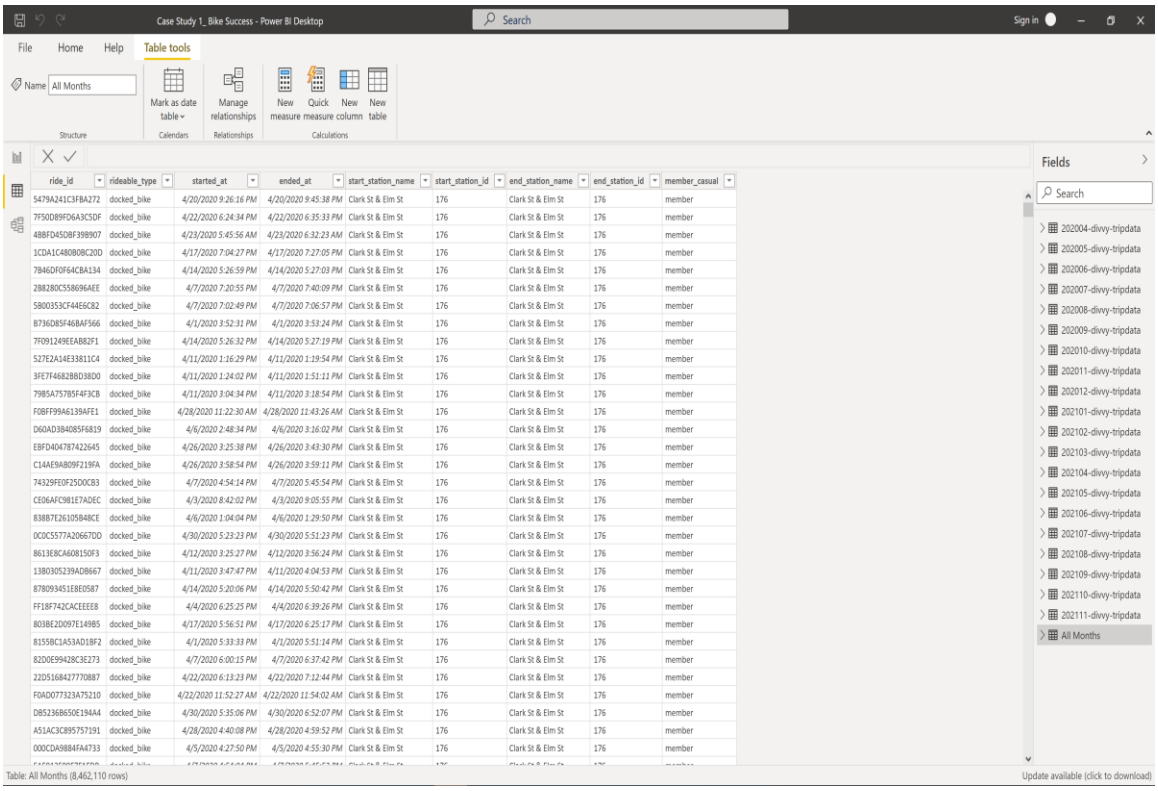
Hence, Insights gained from analyzing this data set should be taken as conclusive and used for business recommendations. The dataset consists of 20 csv files (Apr 2020 to Nov 2021). The columns included (ride_id, rideable_type, started_at, ended_at, start_station_name, start_station_id, end_station_name, end_station_id, start_lat, start_lng, end_lat, end_lng, member_casual).

Phase 3 - Process

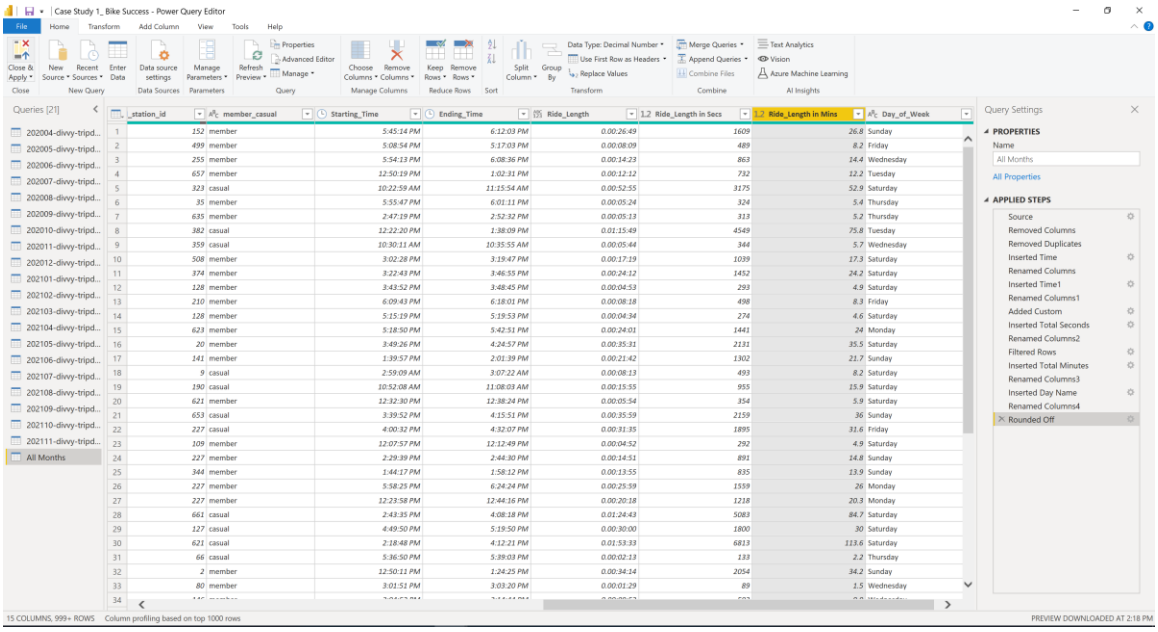
In the Process phase, we clean and validate the data to make sure it's accurate, complete and to make sense of it. To do so, we follow these steps:

1. Explore the data
2. Identify any missing or null values and treat it
3. Make sure the data format is correct

Additionally, we worked with copies of the data and stored the original data on a secured hard drive. We're using Microsoft Power BI for data cleaning, transformation, analysis and visualization. First, we connected to the data set to preview the data and get familiarized with it. Then, we checked the format for all the columns in every table. Afterwards, we appended all rows together from the 20 datasets. Creating one table with 8,462,319 rows. We then removed the duplicates (209) reaching 8,462,110 rows.



Afterwards, we extracted the time component of the (started_at) and (ended_at) columns and created new columns (starting_time) and (ending_time). Then we created another new column called (Ride_Length) to determine the length of the ride by subtracting the (ending_time) from the (starting_time). Then we converted it into 2 more columns (Seconds & Minutes) and removed all the negative and zeroes values which indicates that the (ending time) occurred before or exactly at (starting time). Additionally, a new column was added to extract the Weekday to specify the days were most rides were made.



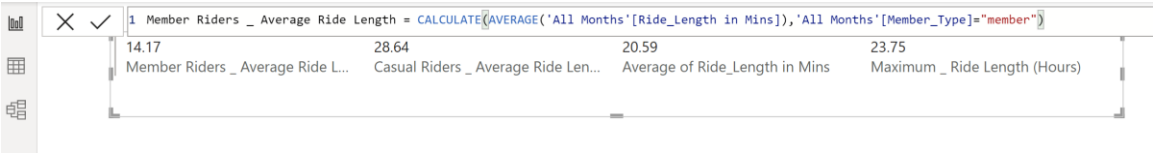
Key takeaways from phase 3:

1- There are 20 main datasets [April 2020 – November 2021]
2- All duplicated rows are removed.
3- All data formats are correct.
4- All incorrect rows are removed

Phase 4 - Analyze

In the analysis phase, we will focus on some basic statistics numbers (Total Number of Rides, Max Ride Length, Mean). Additionally, a list of the top 5 starting & ending stations for both members and casuals were created to identify the locations were most rides start from and end at.

Member_Type	Total Number of Rides
casual	3721963
member	4663529
Total	**8385492**



Top 5 Starting Station Names for Members

Clark St & Elm St	41012
Wells St & Concord Ln	35688
Kingsbury St & Kinzie St	34014
Wells St & Elm St	31944
Dearborn St & Erie St	31459

Top 5 Starting Station Names for Casuals

Streeter Dr & Grand Ave	89049
Millennium Park	49297
Michigan Ave & Oak St	41620
Lake Shore Dr & Monroe St	37467
Theater on the Lake	34978

Top 5 Ending Station Names for Members

Clark St & Elm St	41593
Wells St & Concord Ln	36673
Kingsbury St & Kinzie St	34298
Dearborn St & Erie St	32439
St. Clair St & Erie St	32008

Top 5 Ending Station Names for Casuals

Streeter Dr & Grand Ave	93591
Millennium Park	50919
Michigan Ave & Oak St	43302
Theater on the Lake	38088
Lake Shore Dr & Monroe St	35936

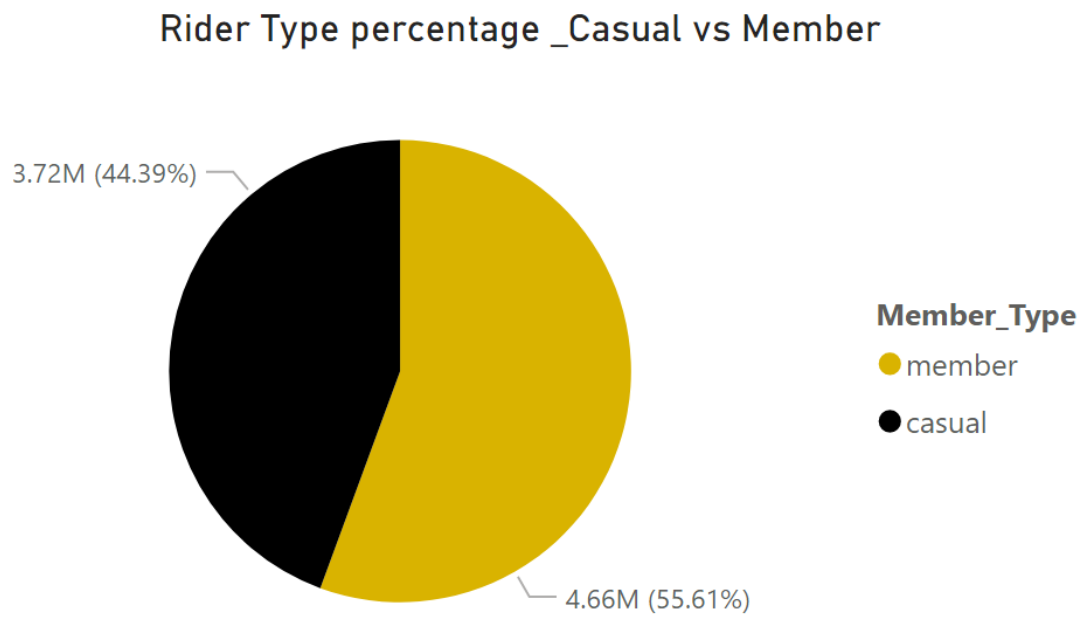
Key takeaways from the statistical summary:

1- The total number of rides is 8,385492 rides where members had the biggest share with 4,663,529 rides while casual riders logged 3,721,963 rides.
2- The maximum ride length was almost 24 hours (probably the rider forgot to log out)
3- The average ride length was 21 minutes. However, casual riders had on average double (28.64 minutes) the trip length than the member riders (14.17 minutes). The reason behind this can be that member riders use their bikes mostly for going to work while casual riders use their bikes for long relaxation rides.

Phase 5 - Share

In the share phase, we will create visualizations to illustrate our findings.

In order to reach our goal of converting casual riders to member riders, we need to know the current situation of the customer base, thus the following pie chart was created to identify the percentage of casual and member riders:

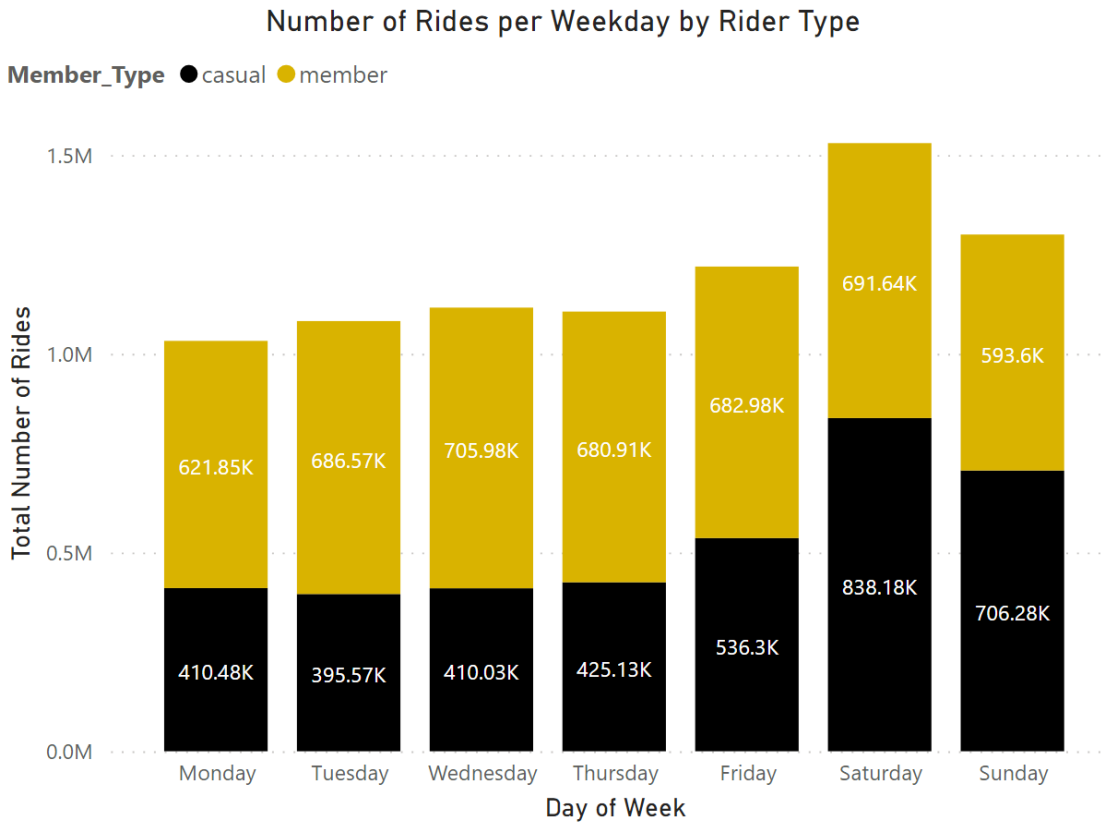


Key takeaways from the pie chart are:

1- Member riders form the biggest share of the customer base (55.6%) while casual riders account for (44.4%)

2- In terms of goal-oriented marketing strategy, the vision of converting casual riders to member riders seems to be accurate as it can be seen from the chart that there is a space for converting more casual riders into member riders.

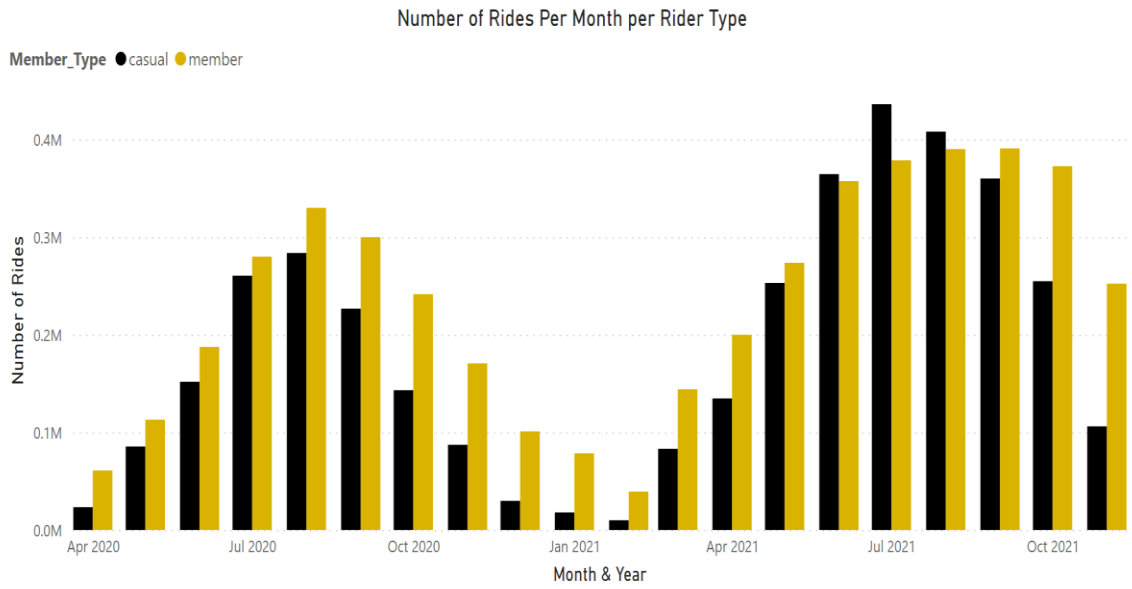
To determine the day of week with most trips for both casuals and members, the following stacked column chart was created:



Key takeaways from the stacked column chart are:

1- For both member and casual riders, Saturday had the most recorded trips.
2- For casual riders, Tuesday was the lowest day with recorded trips
3- For member riders, Monday was the lowest day with recorded trips
4- As an analysis for both riders' behavior, it can be seen that member riders use their biker almost consistently throughout the week. Meanwhile, there is a huge drop in bike usage from weekend (838.18K) to weekday (410.48k) among casual riders. This comes in agreement with the hypothesis we mentioned earlier that casual riders mostly use their bikes for long relaxing rides on the weekends meanwhile member riders use it consistently throughout the week for commuting to work. With this in mind, this can be leveraged during the market strategy planning to convert casual riders into member riders since already some of them use the bikes during the weekdays.

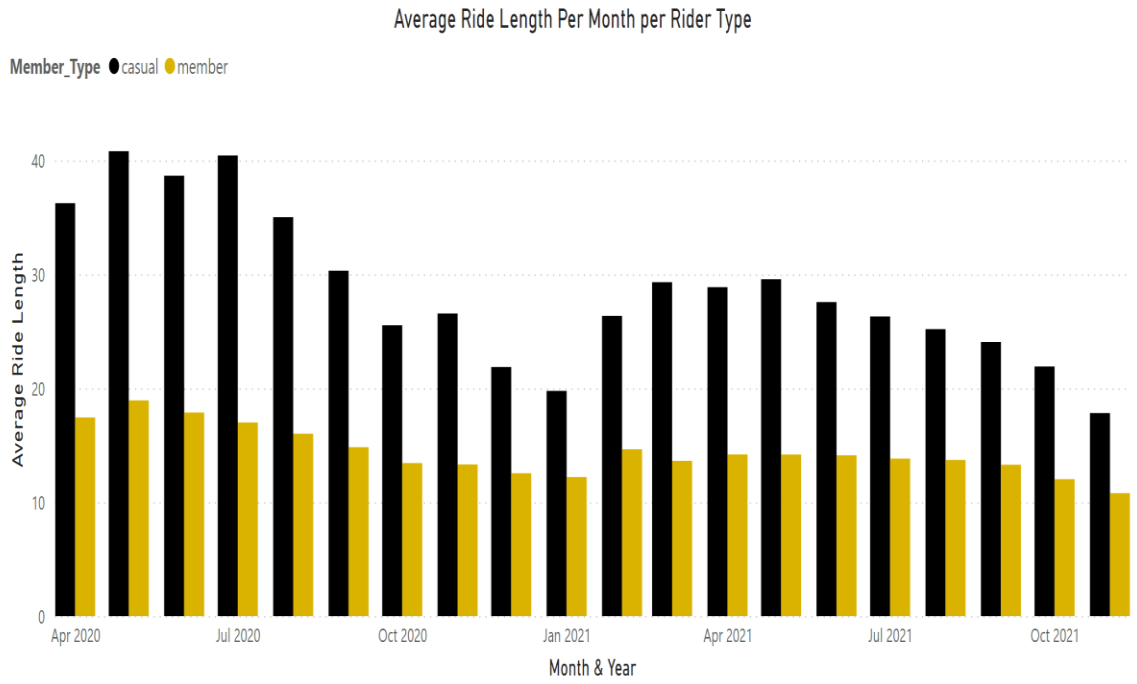
To determine the impact of seasonality on the number of rides taken per month, the following clustered column chart was created:



Key takeaways from the clustered column chart are:

1- The impact of seasonality can be seen clearly since the number of rides during winter is much lower than the number of rides during the other seasons.
2- However, member riders are more likely to withstand the winter season and go on bike trips more than casual riders.
3- The month with the highest number of rides was July 2021 while Feb 2021 recorded the lowest number of rides.

To determine the impact of seasonality on the average length of rides taken per month, the following clustered column chart was created:

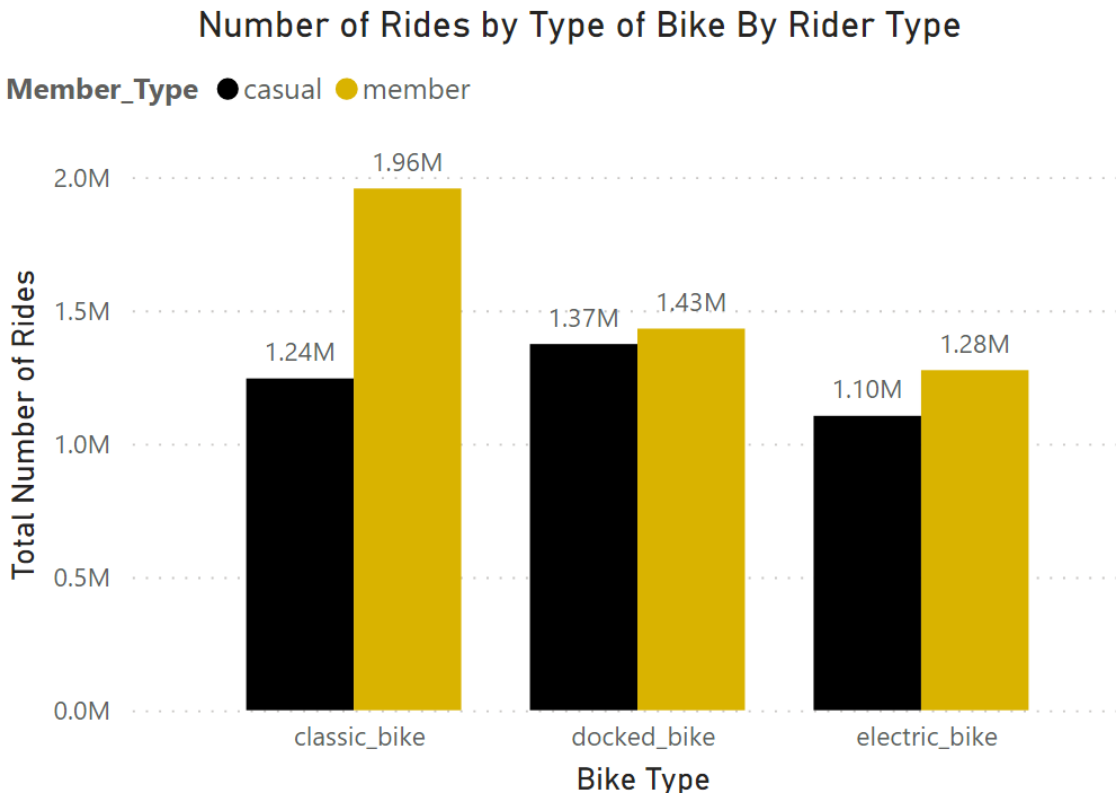


Key takeaways from the clustered column chart are:

1- The expected theory that warmer seasons will have longer rides than colder seasons can be seen clearly in the chart.

2- November 2021 recorded the lowest average ride length (12.88 minutes) while May 2020 had the longest average ride length (28.36 minutes)

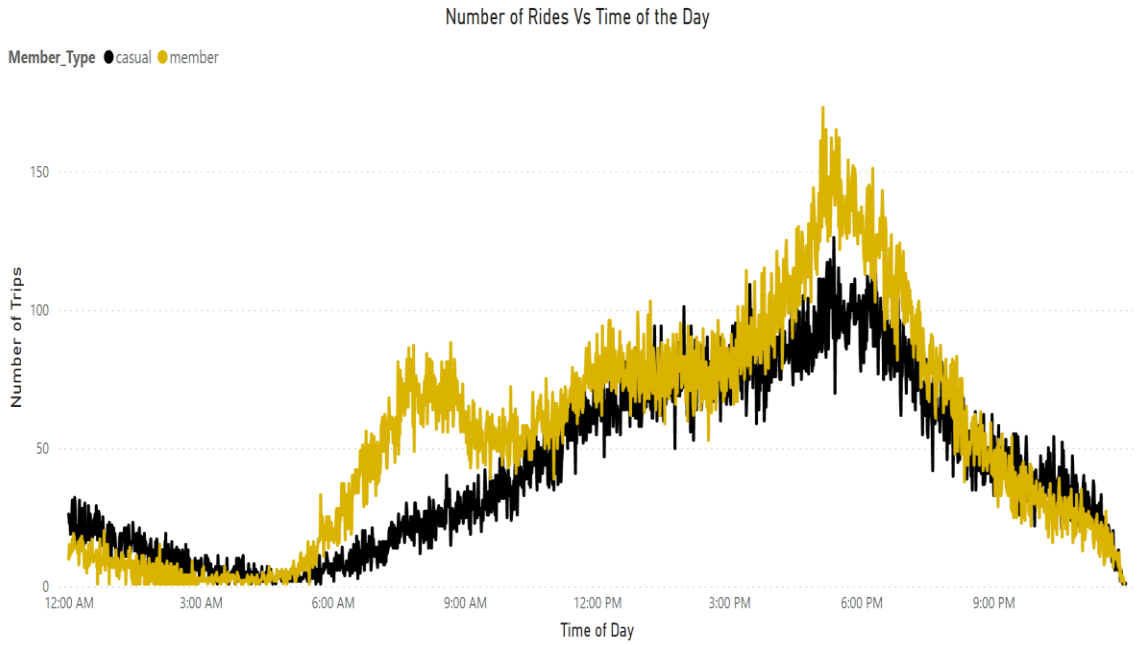
To determine which bike type is mostly used by member and casual riders, the following clustered column chart was created:



Key takeaways from the clustered column chart are:

1- The classic bike is the most used bike with a total of 3,201,157 rides made followed by the docked bike (2,804,700) and lastly the electric bike (2,379,635).
2- For member riders, the classic bike is the favored bike with a total of 1,956,797 rides while the electric bike is the least favored bike with 1,275,456 rides.
3- For casual riders, the docked bike is the favored bike with a total of 1,373,424 rides while the electric bike is the least favored bike with 1104,179 rides.
4- This can be leveraged to convert casual riders to member riders by making and advertising more docked bikes and less electric bikes.

To determine what time of the day, do riders usually make trips, the following line chart was created:



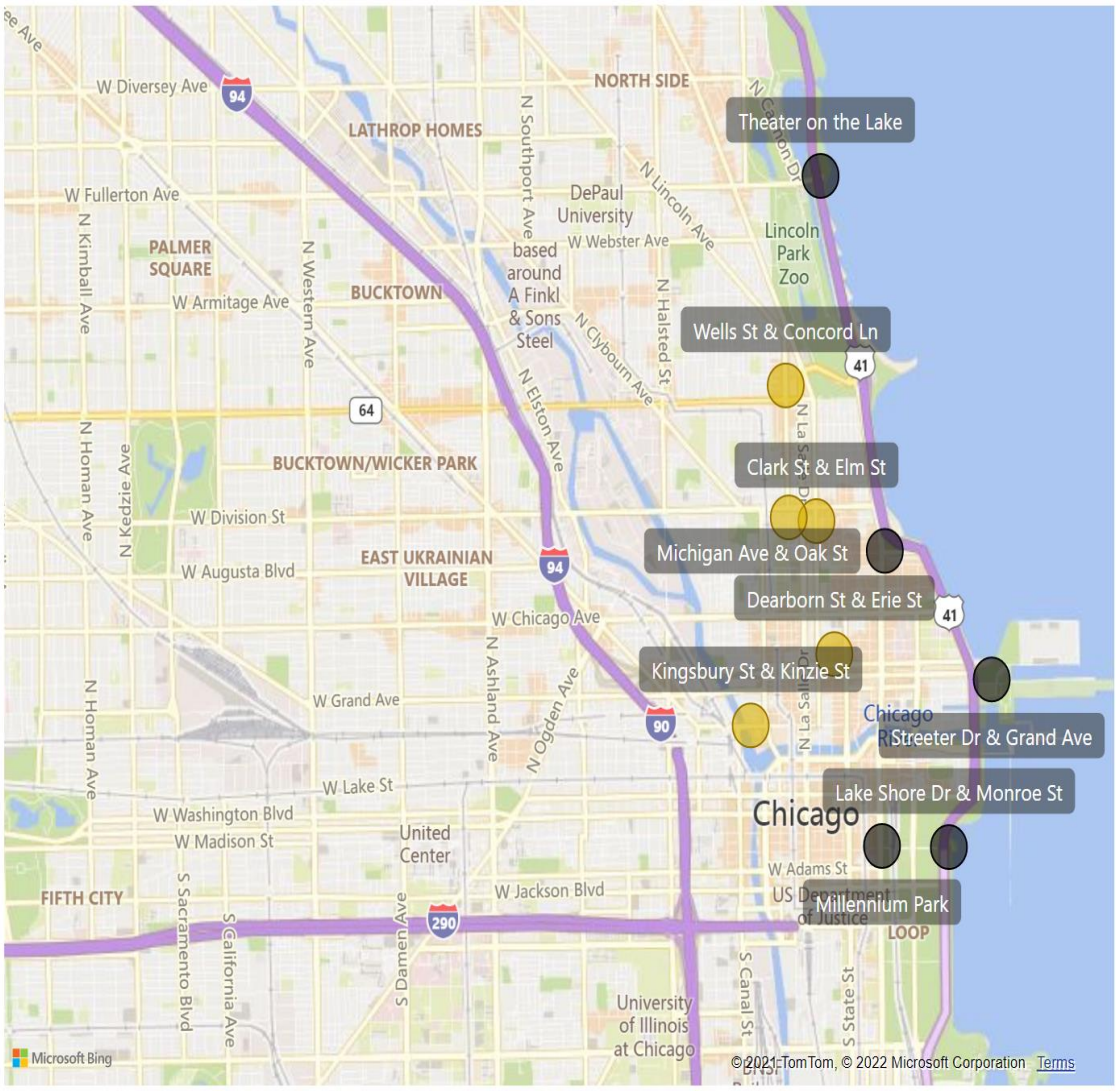
Key takeaways from the line chart are:

1- Casual riders have only one peak which is between 5:00-7:00 PM while member riders have 2 peaks, one between 7:00-9:00 AM and the other is between 5:00-7:00 PM.

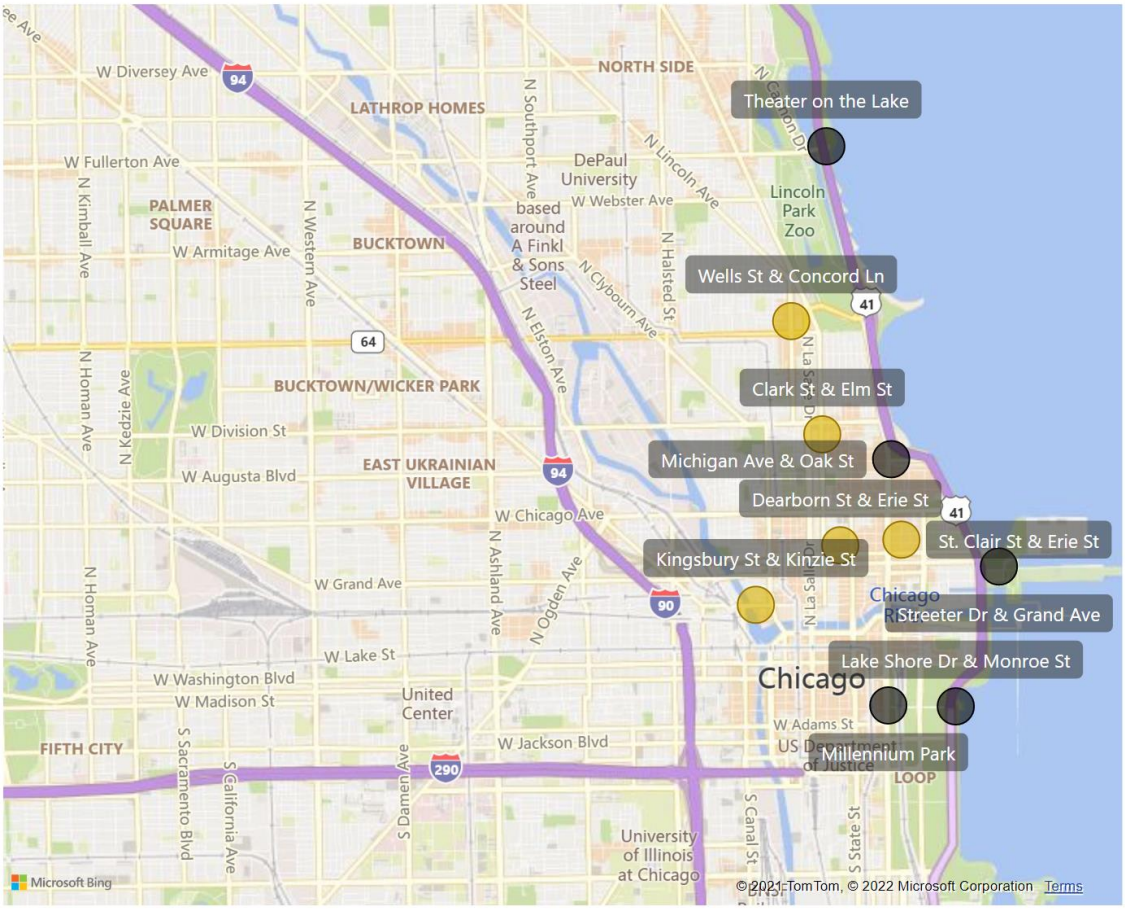
2- This confirms our previous hypothesis that member riders usually use their bikes for commuting to work (7:00-9:00 AM)

To determine the starting and ending locations with most trips for both casual & member riders, the following maps were created:

Top 5 Starting Stations for Members & Casuals



Top 5 Ending Stations for Members & Casuals



Key takeaways from the maps are:

1- Most trips were at the center of the city.
2- Casual riders' stations (Black) were closer to the sea while member riders' stations (Gold) were more at the center of the city.
3- This comes in agreement with our previous hypothesis which stated that casual riders use their bikes for relaxing & leisure purposes. Additionally, Millennium park, one of Chicago most famous parks, is one of the most frequently used stations for casual riders.

Phase 6 - Act

After answering the business question (how do member and casual riders use Cyclistic bikes differently?) previously during the share phase, we will summarize our most important insights, and share our recommendations in this act phase.

Recommendations

1. Due to the impact of seasonality, it is recommended to focus the marketing campaign during the warmer months where most of the rides took place and the majority of riders are willing to go out more frequently.
2. Targeting the busiest stations for casual riders (Streeter Dr & Grand Ave and Millennium Park) for advertisement is a good approach to make sure it reaches the greatest number of riders.
3. According to the established demographic hypothesis that most casual riders use their bikes for leisure rides, it is recommended to advertise the advantages of commuting to work using bikes to convert more casual riders into becoming annual members.
4. In general, promotion offers should be the highest on weekends where most of the riders are outside their home. Additionally, advertisement targeting casual riders should be during their peak hours between 5:00-7:00 PM on weekends while advertisement targeting member riders should be during their peak hours between 7:00-9:00 AM on weekdays and 5:00-7:00 PM throughout the week.
5. Any offers targeting the casual riders should focus on the docked & classic bikes and avoid the electric bikes.
6. To leverage the information that casual riders on average enjoy double the ride length of member riders, it would be wise to offer annual membership incentives based on the duration of ride to convince more casual riders into becoming annual members.
7. A final recommendation is to find a way to track each individual rider (rider_id) in order to study their behavior and understand their patterns. This can help identify casual riders whose riding behavior is similar to member riders, making them a first priority during the marketing campaign.

How Can Bellabeat, A Wellness Technology Company Play It Smart?

Executive Summary

Our case study looks at Bellabeat - a company that produces health-focused smart devices geared towards women. We analyzed 30 Fitbit Fitness users to understand how they use their devices. The discovered insights will then help guide the marketing strategy for Bellabeat.

Phase 1- Ask

At this stage of the data analysis, key stakeholders are identified and questions are posed to guide the process.

Introduction

Urška Sršen and Sando Mur founded Bellabeat, a high-tech company that manufactures health-focused smart products. Sršen used her background as an artist to develop beautifully designed technology that informs and inspires women around the world. Collecting data on activity, sleep, stress, and reproductive health has allowed Bellabeat to empower women with knowledge about their own health and habits. Since it was founded in 2013, Bellabeat has grown rapidly and quickly positioned itself as a tech-driven wellness company for women.

Products

1. Bellabeat app: The Bellabeat app provides users with health data related to their activity, sleep, stress, menstrual cycle, and mindfulness habits. This data can help users better understand their current habits and make healthy decisions. The Bellabeat app connects to their line of smart wellness products.
2. Leaf: Bellabeat's classic wellness tracker can be worn as a bracelet, necklace, or clip. The Leaf tracker connects to the Bellabeat app to track activity, sleep, and stress.
3. Time: This wellness watch combines the timeless look of a classic timepiece with smart technology to track user activity, sleep, and stress. The Time watch connects to the Bellabeat app to provide you with insights into your daily wellness.
4. Spring: This is a water bottle that tracks daily water intake using smart technology to ensure that you are appropriately hydrated throughout the day. The Spring bottle connects to the Bellabeat app to track your hydration levels.
5. Bellabeat membership: Bellabeat also offers a subscription-based membership program for users. Membership gives users 24/7 access to fully personalized guidance on nutrition, activity, sleep, health and beauty, and mindfulness based on their lifestyle and goal

The Stakeholders

1. Urška Sršen: Bellabeat's cofounder and Chief Creative Officer.
2. Sando Mur: Mathematician and Bellabeat's cofounder; key member of the Bellabeat executive team.
3. Bellabeat marketing analytics team: A team of data analysts responsible for collecting, analyzing, and reporting data that helps guide Bellabeat's marketing strategy.

Business Task

In this case study, we will analyze the usage of smart devices to gain insight into how consumers use non-Bellabeat devices. This data will help us determine the best marketing strategy for Bellabeat.

Business Questions to Answer

1. What are some trends in smart device usage?
2. How could these trends apply to Bellabeat customers?
3. How could these trends help influence Bellabeat marketing strategy?

Phase 2 - Prepare

In the Prepare phase, we identify the data being used and its limitations.

Data Source

The Data set is publicly available on Kaggle (FitBit Fitness Tracker Data) and stored in 18 csv files. It contains personal fitness tracker from thirty FitBit users. Thirty eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. It includes information about daily activity, steps, and heart rate that can be used to explore users' habits

Data Limitations

The data is outdated as it was collected in 2016. Thus, users' daily activity and habits may have changed since then. Additionally, the sample size of 30 FitBit users is not representative of the whole FitBit population (More than 30 million worldwide).

Data Quality

A good data set is usually reliable, original, comprehensive, current and cited. However, in our case, the data set is considered to be of a bad quality due to the following reasons:

1. It is not reliable: it has only 30 respondents
2. It is not original: 3rd party provider
3. Somewhat comprehensive: The features of the 2 systems are similar
4. It is not current: Outdated and might be irrelevant
5. It is not properly cited: Collected from a 3rd party

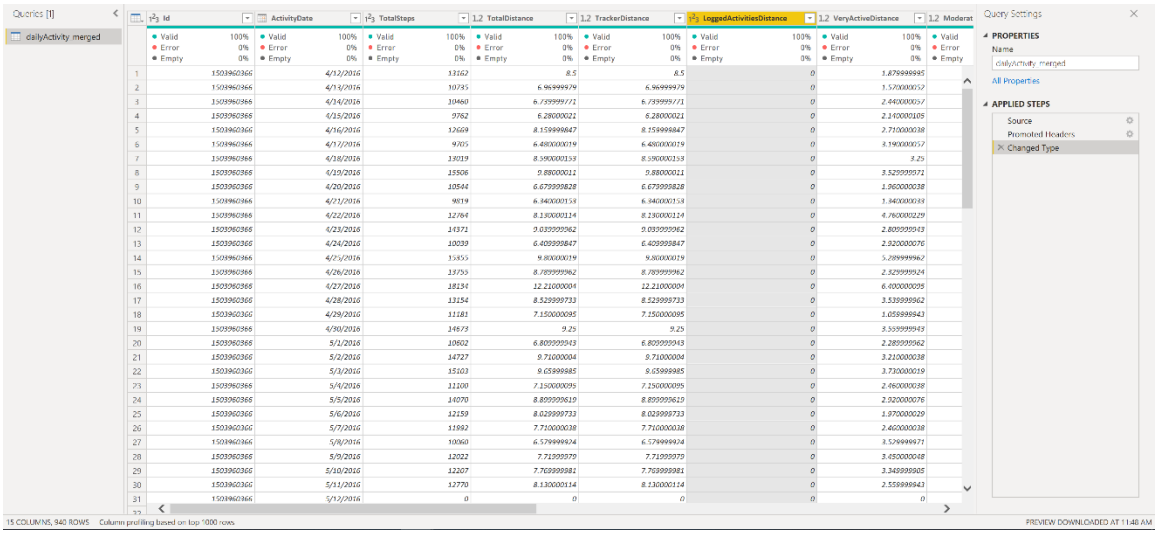
Hence, Insights gained from analyzing this data set should not be taken as conclusive or used for business recommendations. The dataset consists of 18 csv files in total, 15 of them in long format and 3 in wide format. The timeframes for the data collection are second, minute, hour, day, and it focuses on multiple factors, including activity, sleep, weight, calories burned, heart rate, etc. We will focus on daily activity data, weight, sleep, and heart rate data.

Phase 3 - Process

In the Process phase, we clean and validate the data to make sure it's accurate, complete and to make sense of it. To do so, we follow these steps:

1. Explore the data
2. Identify any missing or null values and treat it
3. Make sure the data format is correct

Additionally, we worked with copies of the data and stored the original data on a secured hard drive. We're using Microsoft Power BI for data cleaning, transformation, analysis and visualization. First, we connected to the data set to preview the data and get familiarized with it. Then, we adjusted the format for all the columns in every table. Afterwards, we checked for any null values in the data set using the "Column Quality" option.

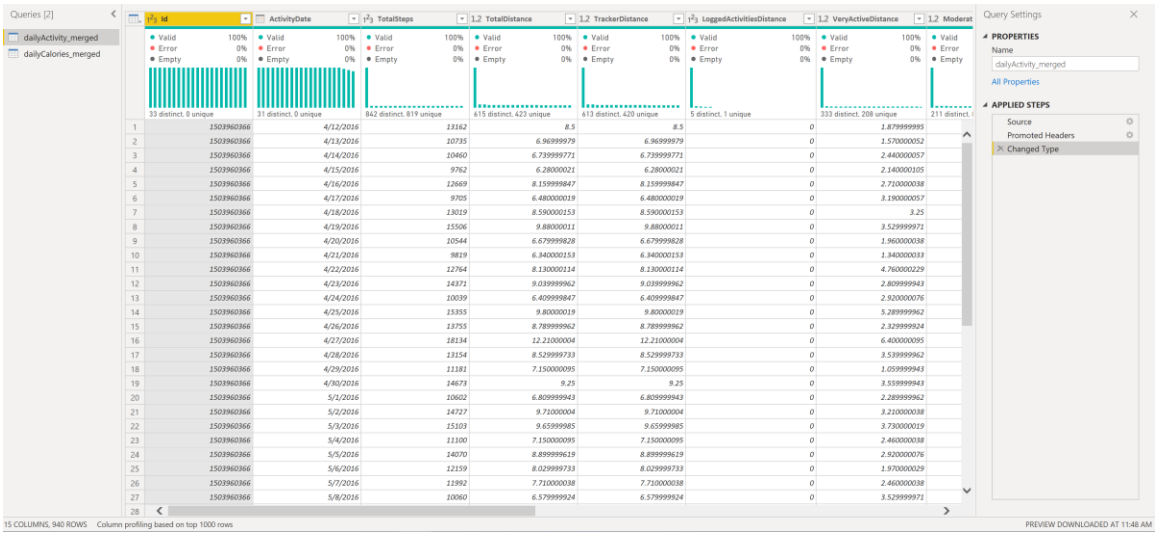


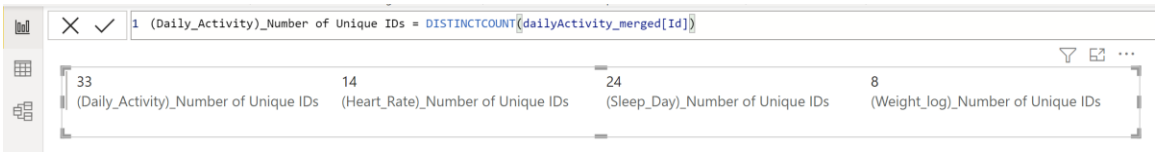
Key takeaways from the exploration and data cleaning:

1. There are 4 main datasets [daily_activity, heart_rate, sleep_day, weight_log]
2. There are no null or missing values in the datasets.
3. All data formats are correct.

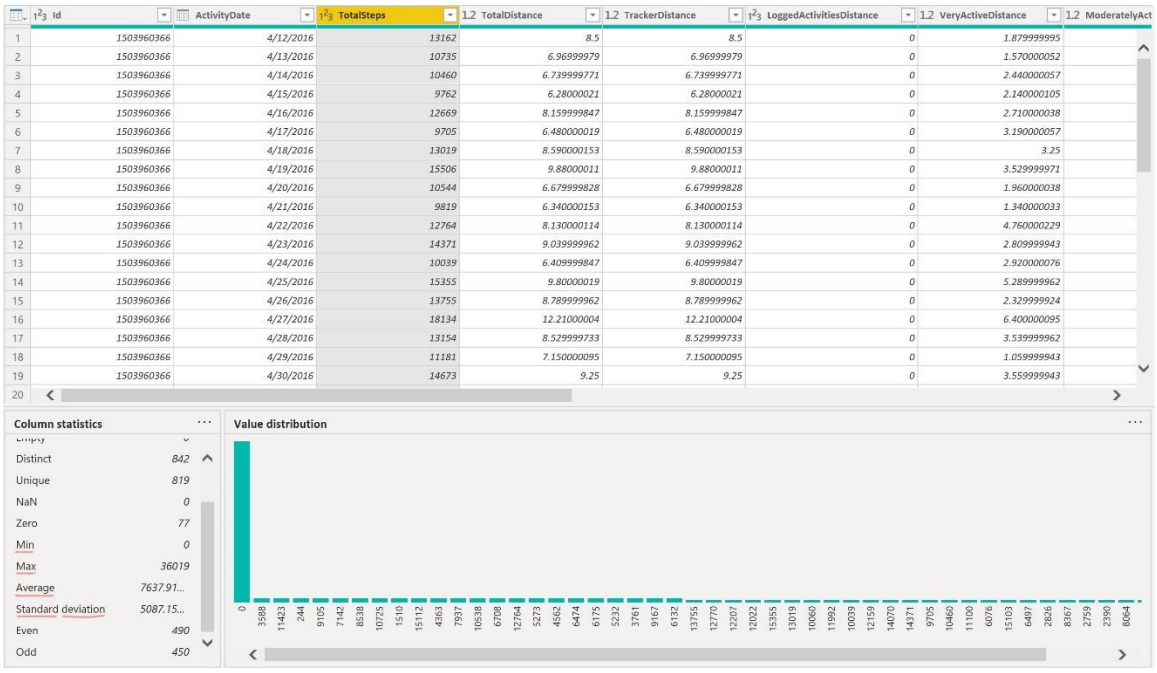
Phase 4 - Analyze

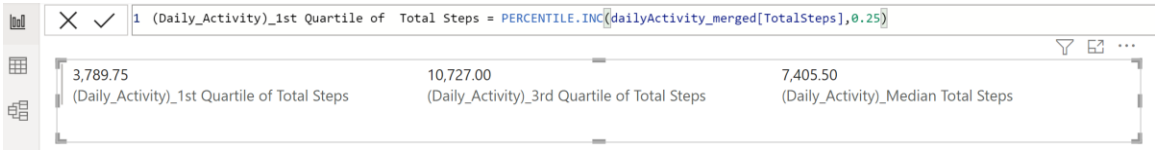
In the analysis phase, we will start by counting the number of unique IDs in every dataset to identify how many unique users have provided their data. This can be done using the “Column Distribution” option or we can create a measure to count the unique values in every data set as shown below:





This shows that the daily_activity data set had the most participants (33), followed by the sleep_day dataset (24), then the heart_rate dataset (14) and lastly the weight_log dataset (8). Next, we will focus on data profiling and showing some basic statistics numbers (Min, Max, Mean, Standard Deviation) for the columns by utilizing the “Column Profile” option. Additionally, a couple of measure were created to identify the 1st percentile, 3rd percentile and the median.





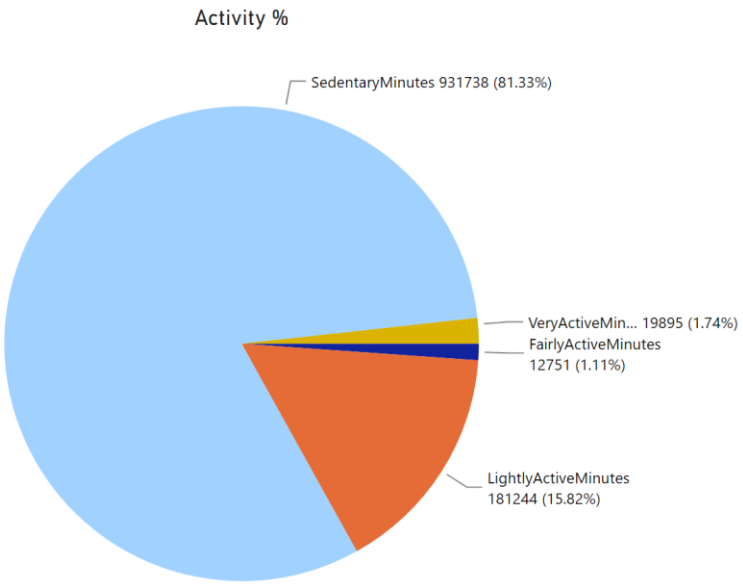
Key takeaways from data profiling & statistical summary:

1. Fitbit users logged an average of 7637 steps daily which is less than the recommend number (10,000) by the Centers for Disease Control and Prevention (CDC).
2. Fitbit users slept 7 hours per day on average
3. Fitbit users have an average BMI of 25.1 which is just outside the normal/healthy range (18.5-24.9) according to the Centers for Disease Control and Prevention (CDC).
4. Fitbit users heart rate ranges from 36 up to 200 where the normal pulse rate ranges from 60-100 for adults according to American Heart Association (AHA).
5. Light users recorded an average logging of 192.81 minutes (16%), fairly users recorded an average logging of 13.56 minutes (1%) while sedentary users recorded an average logging of 991.2 minutes (83%).

Phase 5 - Share

In the share phase, we will create visualizations to illustrate our findings.

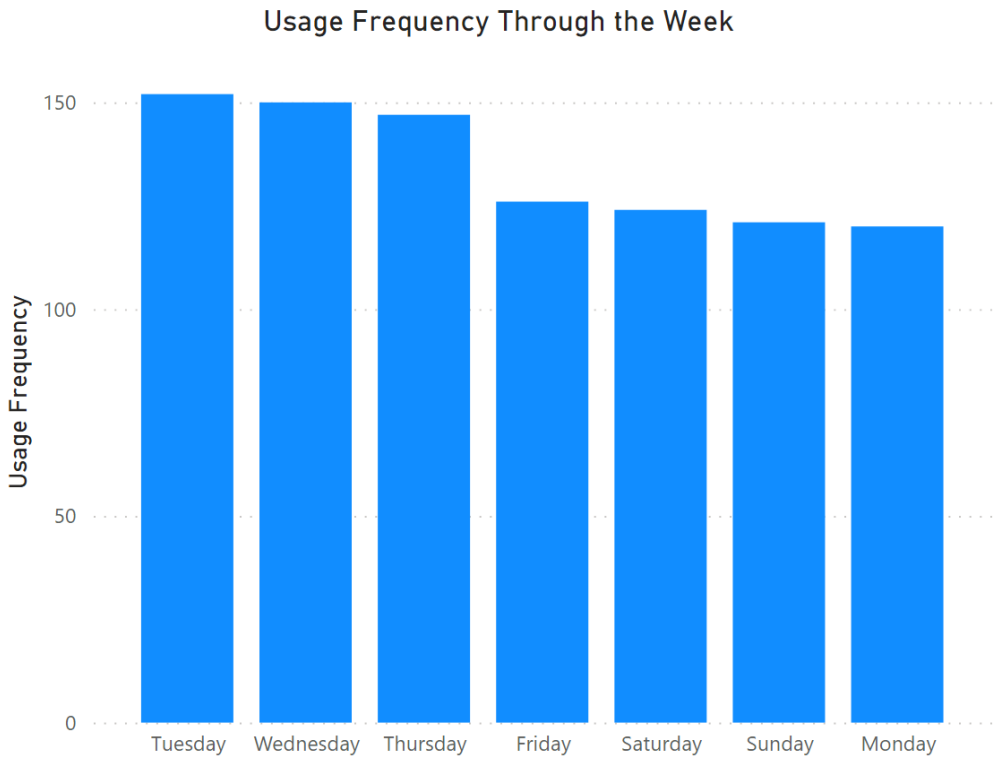
To determine the percentage of each type of activity, the following pie chart was created:



Key takeaways from the pie chart are:

1. The fairly active has the lowest % (1.11) while the sedentary has the highest % (81.3)
2. This indicates that the fitness activities (fairly active or very active) are rarely tracked by the app (1.11 % and 1.74%), respectively.
3. However, the app is mostly used to track the sedentary activities (81.3%) such as commute to work, shopping, moving from one room to another.
4. It can be deduced that the app is not used for the main purpose it was designed for which is to track the fitness activities.

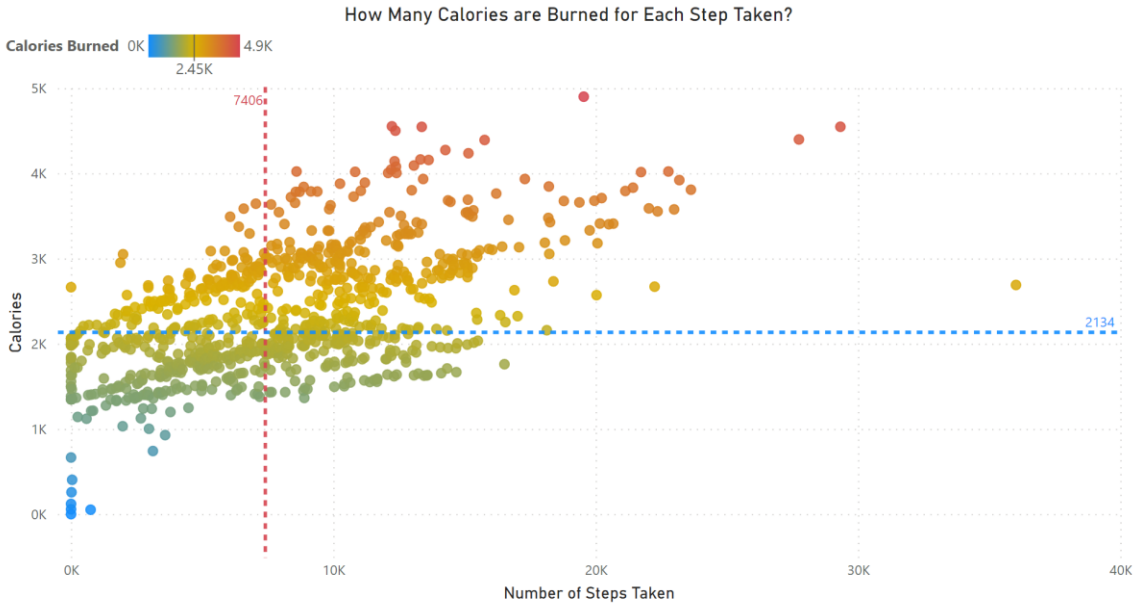
To determine the usage frequency, the following column chart was created to show the number of time users logged into the Fitbit app throughout the week:



Key takeaways from the column chart are:

1. The usage frequency graph shows that the users logged into the application more frequently during the middle of the week until Friday.
2. However, the number of times the users logged into the app have dropped from 152 on Tuesday, to 126 on Friday till reaching 120 on Monday.
3. This might indicate that they tend to care less about the app during the weekends.

To determine the relationship between the number of steps taken and the calories burned, the following scatter plot was created:



Key takeaways from the scatter plot are:

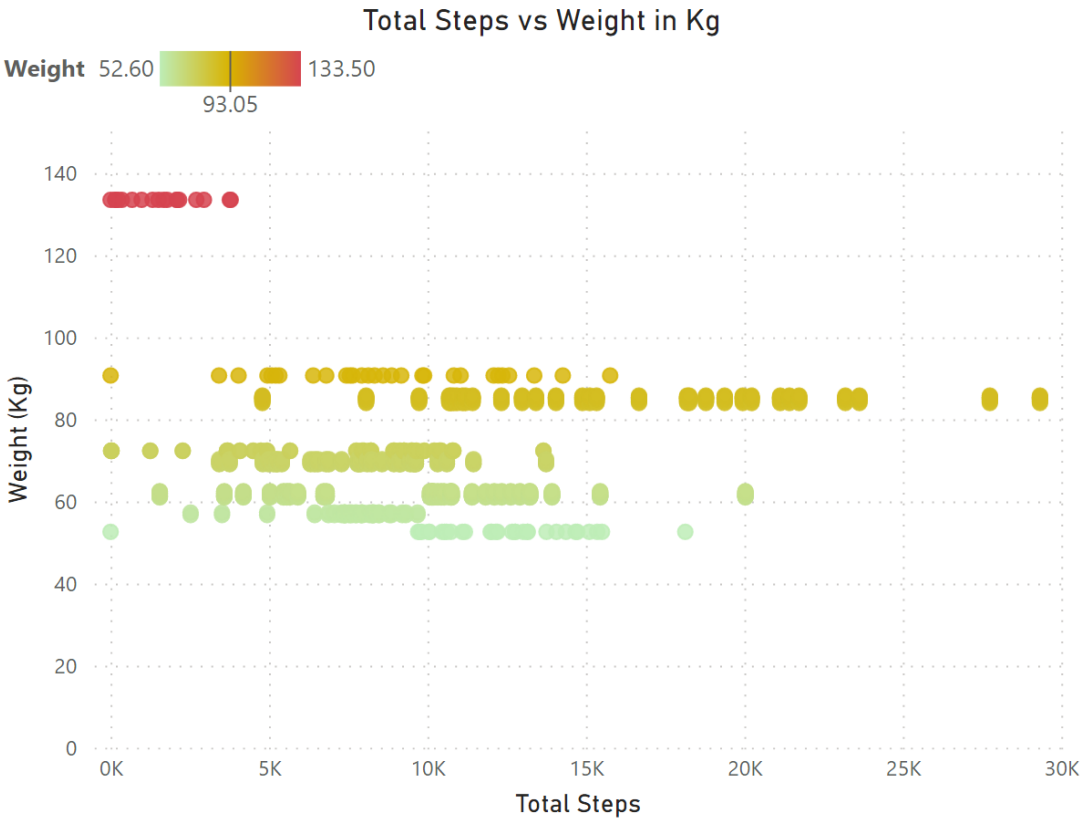
1. There is a strong positive correlation which means that as the number of steps taken increase, the number of calories burned increases as well.
2. An outlier was noticed where the number of steps were 36019 while the number of burned calories were 2690. This could be caused by error in data calculation by the application.
3. To avoid skewing the results, the Median of the number of steps taken (7406) and the number of calories burned (2134) was plotted instead of the average.
4. The bulk of the calories burned was on the left part of the x-axis ranging from 0 to 15000 steps.

To determine the relation between the number of hours logged into the app and the number of calories burned, the following scatter plot was created:

Key takeaways from the scatter plot are:

1. There is a weak positive correlation which means that as the number of hours logged into the app increase, it won't necessarily lead to an increase in the number of calories burned. This can be due to the median high sedentary hours (18) as it can be seen on the plot.
2. An outlier was noticed where the number of hours logged were 24 while the number of burned calories were zero. This could be caused by error in data calculation by the application.
3. To avoid skewing the results, the Median of the sedentary hours (18) and the number of calories burned (2134) was plotted instead of the average.

To determine the relationship between the total steps taken and the weight, the following scatter plot was created:



Key takeaways from the scatter plot are:

1. Users with higher weights tend to take less steps as it can be seen by the red bubbles
2. However, the majority of the data shows that the weight can remain constant even with higher number of steps taken. This can be due to unhealthy eating habits.

Phase 6 - Act

In the act phase, we will summarize our most important insights, answer the business questions that were drafted during phase 1 and share our recommendations.

Business Questions

1. What are some trends in smart device usage?
    a. Users tend to track their activities less on weekends and more on weekdays
    b. Users rarely use the app to track fitness or healthy activities (1.1 % for fairly active activity & 1.7% for very active activity). However, most of them use it to track daily activities (81.3% for sedentary activity).
2. How could these trends apply to Bellabeat customers?
    a. They both offer products that aim to provide women with their health, habit and fitness data in order to help them make better choices regarding their health. These common trends about health and fitness can certainly be applied to Bellabeat products.
3. How could these trends help influence Bellabeat marketing strategy?
    a. Bellabeat app should send a notification message on weekends to motivate users to workout.
    b. Each user can choose a specific goal/plan when first starting the app, then the app can help him/her achieve it by providing data about how many calories does he need to burn, how many steps need to be taken to accomplish his daily goal, the optimum daily calorie intake and easy-to-make recipes for healthy meals.
    c. Bellabeat app should also monitor the health rate of their users and notify them if it went too high or too low. Thus, helping them to take quick action and save their lives.
    d. A monthly report should be sent to the users detailing the number of calories they burned during that month, a graph showing their heart rate and highlighting any abnormal values, the total number of steps taken and how it is correlated to the user weight-loss journey.

Testing Hypothesis: Seasonal Variation of Electricity and Water Consumption is different

Understand the Business Case

- Business outcomes to impact (sector targeting for Water & Electricity conservation campaigns)
- Key Stake holder (power stations, desalination plants, conservation department)
- Fitting into the business strategy (demand forecasting, public awareness, customer behavior)

Measurement Plan

- Define KPIs (Amount of sold power, Amount of sold water)
- Data required to track KPIs (Monthly Water & Electricity Consumption (2018-2020), Water & Electricity Consumption by sector (2018-2020))

Data Collection & Preparation

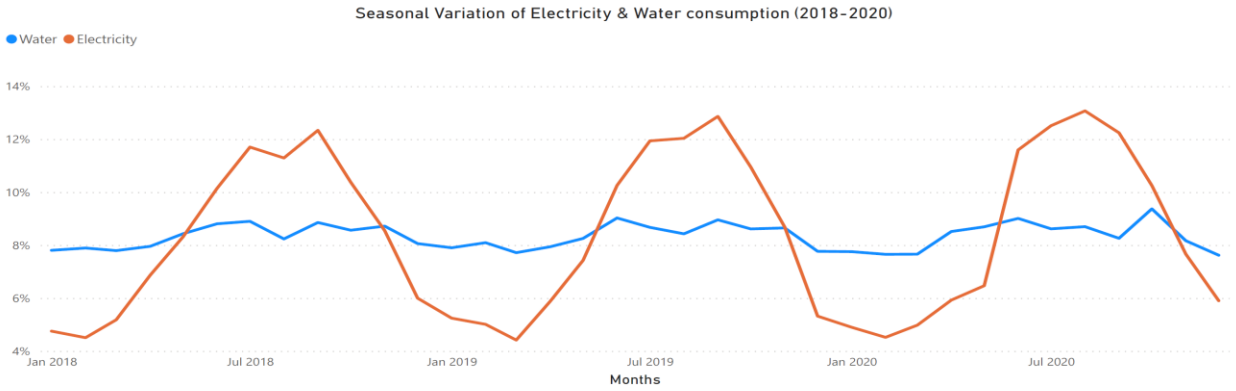
- Data was collected from the Research & Studies annual statistical book
- Quality Assurance (QA) and data profiling was performed to make sure the data was clean.

Understand the Data

- Representation of each record (Each row represented a monthly water & electricity consumption
- Relevant fields (Total Sold Water, Total Sold Electricity)

Analyze & Visualize

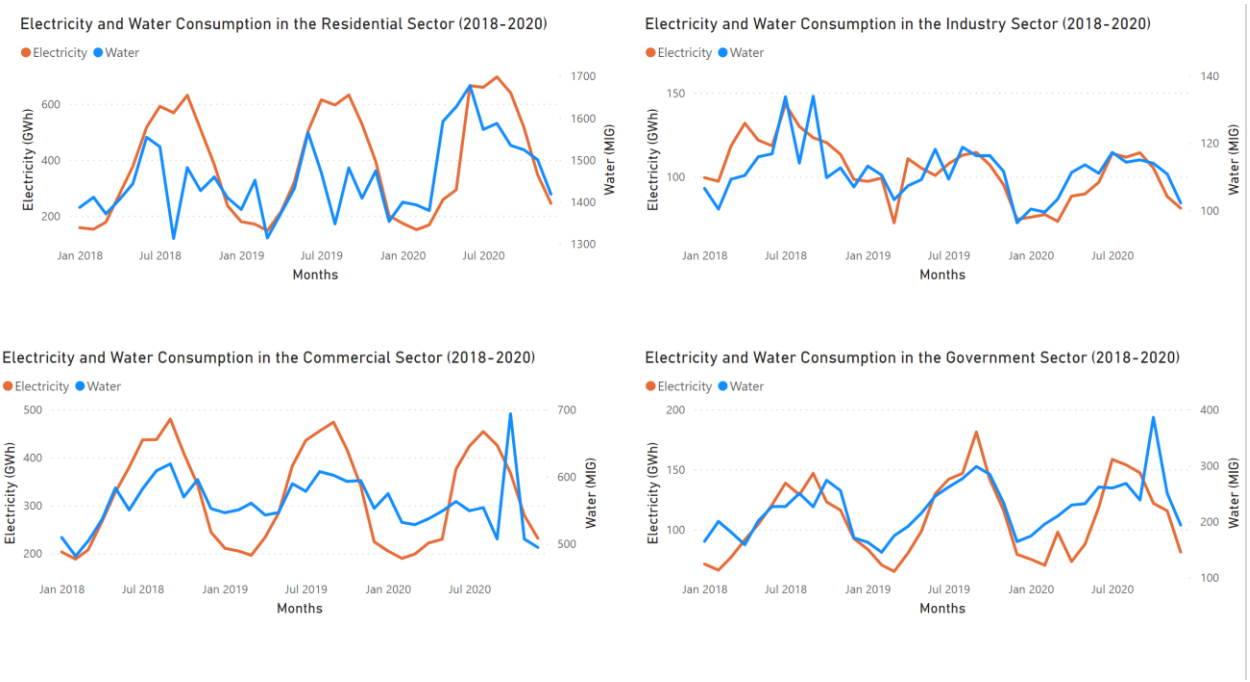
After collecting the data in Excel, the file was exported to Power BI to start the analysis. A number of measures were created to determine the percentage of monthly sold water and electricity from 2018 to 2020. The following graph was generated:

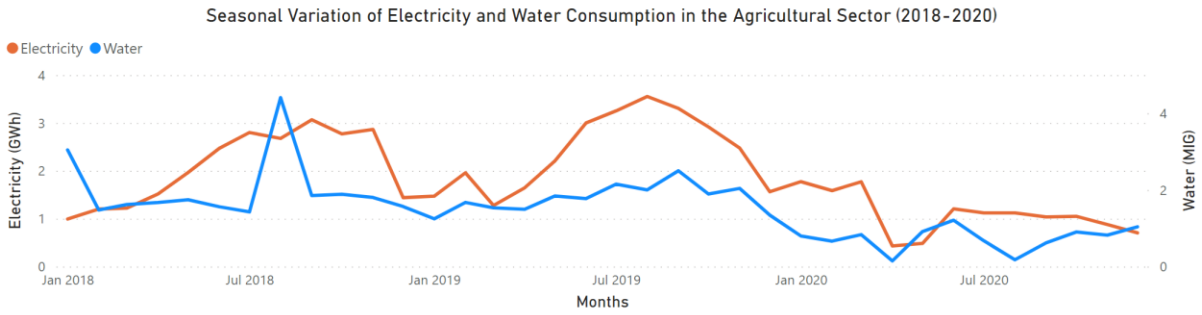


Electricity			Water		
144.76% (2018)	151.17% (2019)	162.82% (2020)	10.65% (2018)	9.89% (2019)	10.87% 2020

The assumptions used during the analysis were that the summer season included July, August and September (Q3) and the winter season included January, February and March (Q1). As it can be seen from the graph, the seasonal variation in electricity consumption is much higher than in water consumption. The difference between the electricity consumption in summer and winter is 144.76% in 2018, 151.17% in 2019 and 162.82% in 2020. Meanwhile, the difference between the water consumption in summer and winter is 10.65% in 2018, 9.89% in 2019 and 10.87% in 2020.

Afterwards, we decided to dive deeper and determine the seasonal variation for electricity and water consumption by sectors. Additional measures were created and the following charts were generated:

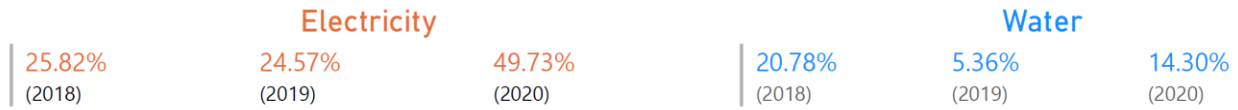




For the **residential** sector, the seasonal variation in electricity and water consumption were as follows:

Electricity			Water		
266.45%	270.75%	305.40%	3.75%	3.63%	12.52%
(2018)	(2019)	(2020)	(2018)	(2019)	(2020)

For the **Industrial** sector, the seasonal variation in electricity and water consumption were as follows:



For the **Commercial** sector, the seasonal variation in electricity and water consumption were as follows:

For the **Government** sector, the seasonal variation in electricity and water consumption were as follows:

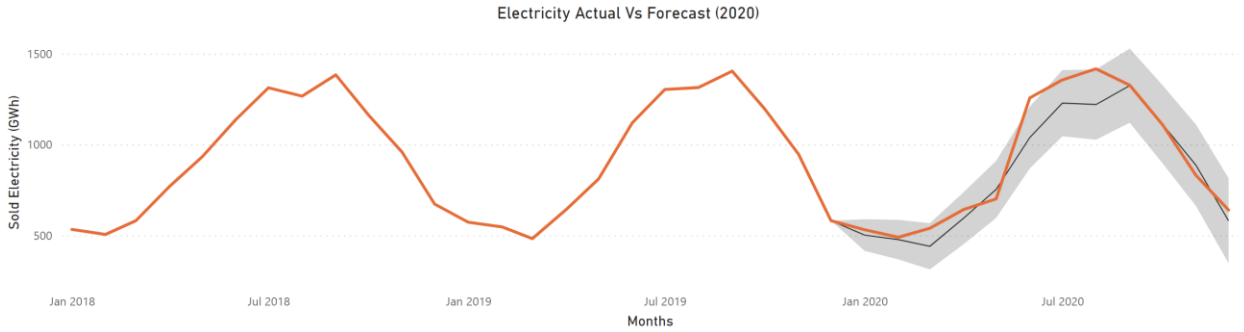
For the **Agricultural** sector, the seasonal variation in electricity and water consumption were as follows:

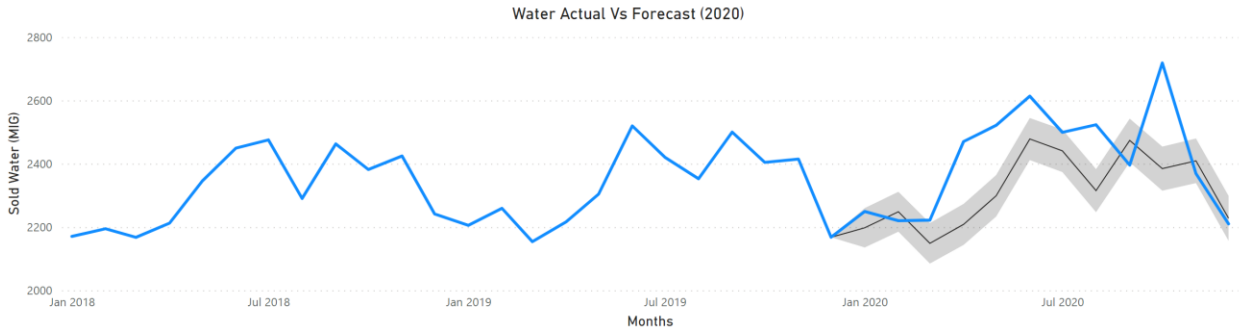
The following conclusions were drawn:

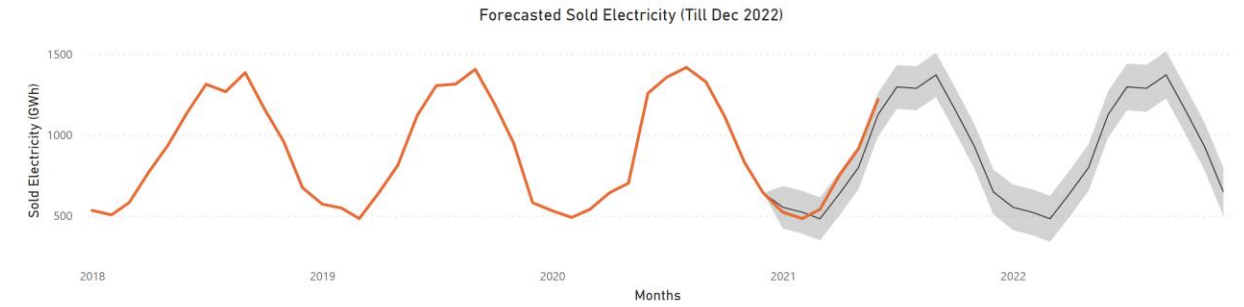
- The Residential sector had the highest seasonal variation in electricity consumption and the lowest seasonal variation in water consumption except for the year 2020 (Covid-19 impact).
- The industrial sector had the lowest seasonal variation in electricity consumption. This means that the factories electricity consumption is almost consistent during Summer and Winter.
- In 2020, the water consumption in the commercial sector was higher in winter (before lockdown) than in summer (After lockdown).
- The government sector electricity and water consumption had inconsistent patterns throughout the years.
- The agricultural sector electricity and water consumption showed anomalies in the year 2020.

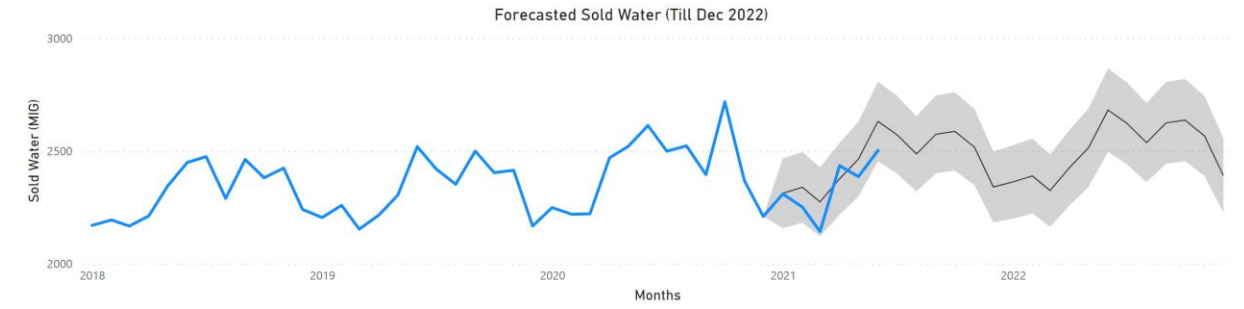
Data-Driven Insights

- The residential sector should be targeted by electricity conservation campaigns.
- The government sector should be further investigated for the issue of unbilled water and electricity meters.
- The agricultural sector should be further investigated by the concerned departments to determine the root causes of the anomalies in the year 2020.
- Additionally, the time series forecasting technique was utilized in Power BI to compare the actual water and electricity consumption in 2020 and the forecasted consumption in order to determine the accuracy of the forecasting model. Afterwards, the technique was used, along seasonality, to predict the electricity and water consumption till December 2022. This can act as a guide line for water and electricity demand forecasting.


Electricity Actual Vs Forecast (2020)
Sold Electricity (GWh)
1500
1000
500
Jan 2018
Jul 2018
Jan 2019
Jul 2019
Jan 2020
Jul 2020
Months


Water Actual Vs Forecast (2020)
Sold Water (MIG)
2800
2600
2400
2200
2000
Jan 2018
Jul 2018
Jan 2019
Jul 2019
Jan 2020
Jul 2020
Months


Forecasted Sold Electricity (Till Dec 2022)
Sold Electricity (GWh)
1500
1000
500
2018
2019
2020
2021
2022
Months


Forecasted Sold Water (Till Dec 2022)
Sold Water (MIG)
3000
2500
2000
2018
2019
2020
2021
2022
Months

How can the sold power in Autumn be greater than the generated power?

Understand the Business Case

- Business outcomes to impact (Power plants efficiency, Revenue, Losses)
- Key Stake holder (Electricity generation, transmission and distribution departments)
- Fitting into the business strategy (Increase revenue, Decrease losses)

Measurement Plan

- Define KPIs (Amount of power generated, Amount of power sold, Losses in the network, Measurement tools efficiency)
- Data required to track KPIs (Manual readings, Billing dates, Drive-by readings)

Data Collection & Preparation

- Manual readings were collected from the billing system (CC&B)
- Automatic readings were collected via drive-b
- Quality Assurance (QA) and data profiling was performed to clean the data and remove any null or duplicates instances.

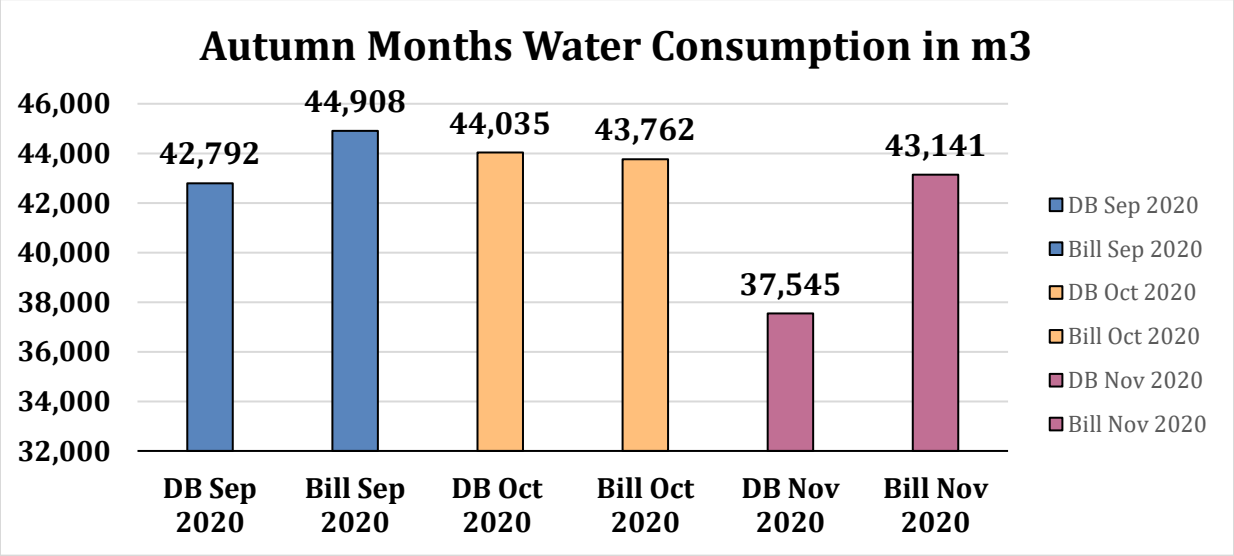
Understand the Data

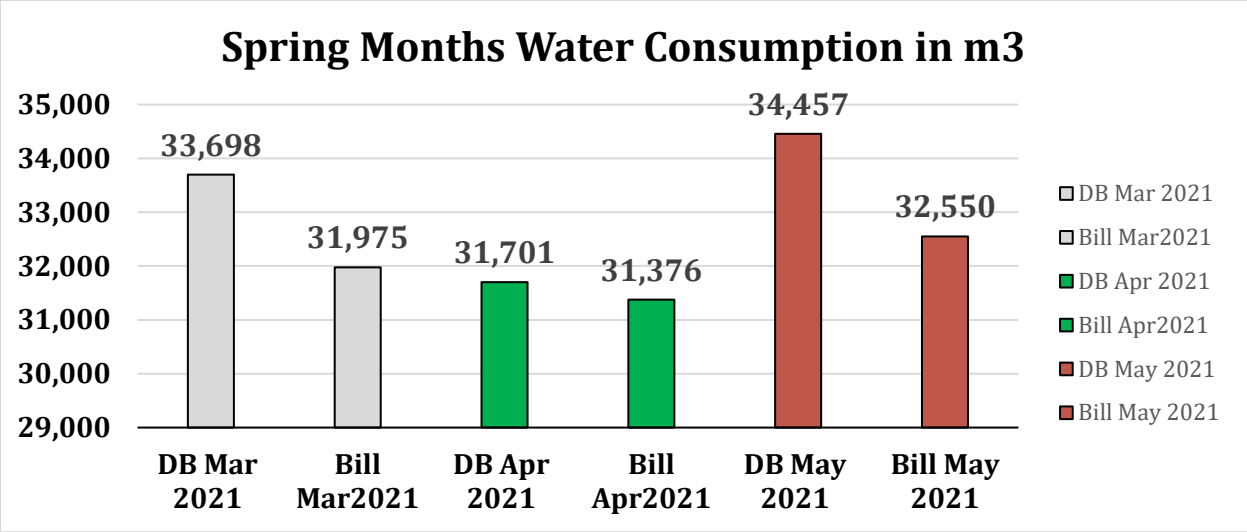
- Representation of each record (Each row represented monthly readings for a single water meter)
- Relevant fields (Consumption, date of reading)

Analyze & Visualize

After conducting a 6 months comparison between the drive-by reading and manual bill reading of 1500 smart water meters in Maysaloon area, the following outcomes were found:

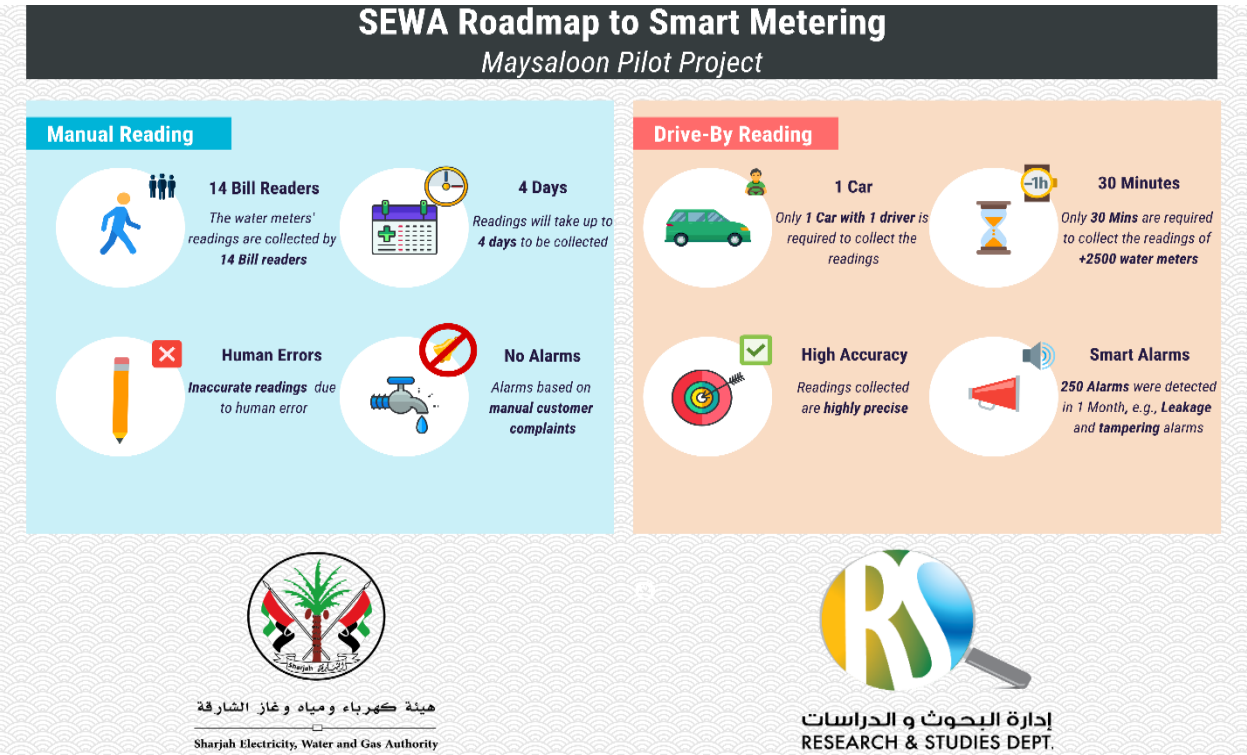
- In autumn, for the months (September and November), the billing consumption was higher than the drive-by consumption and almost equal in October.
- In spring, for the months (March, April and May), the drive-by consumption was higher than the billing consumption.
- This highlights the impact of billing cycle on the readings. For instance, in autumn, the billing is higher because in September bill, the billing cycle reading is taken from **15/8** to **15/9** which means 15 additional hot days (second half of August) with high consumption are added to September bill instead of it being from 1/9 to 30/9.
- Similarly, in spring, the billing is lower because in May bill, the billing cycle reading is taken from **15/4** to **15/5** which means 15 fewer hot days (second half of May) with high consumption are removed from May bill instead of it being from 1/5 to 31/5.
- This can explains & justifies why in the **Emirate of Sharjah Generated, sent out and sold Electrical Power** graph seen below, the sold power in September, October and November is higher than the generated power.





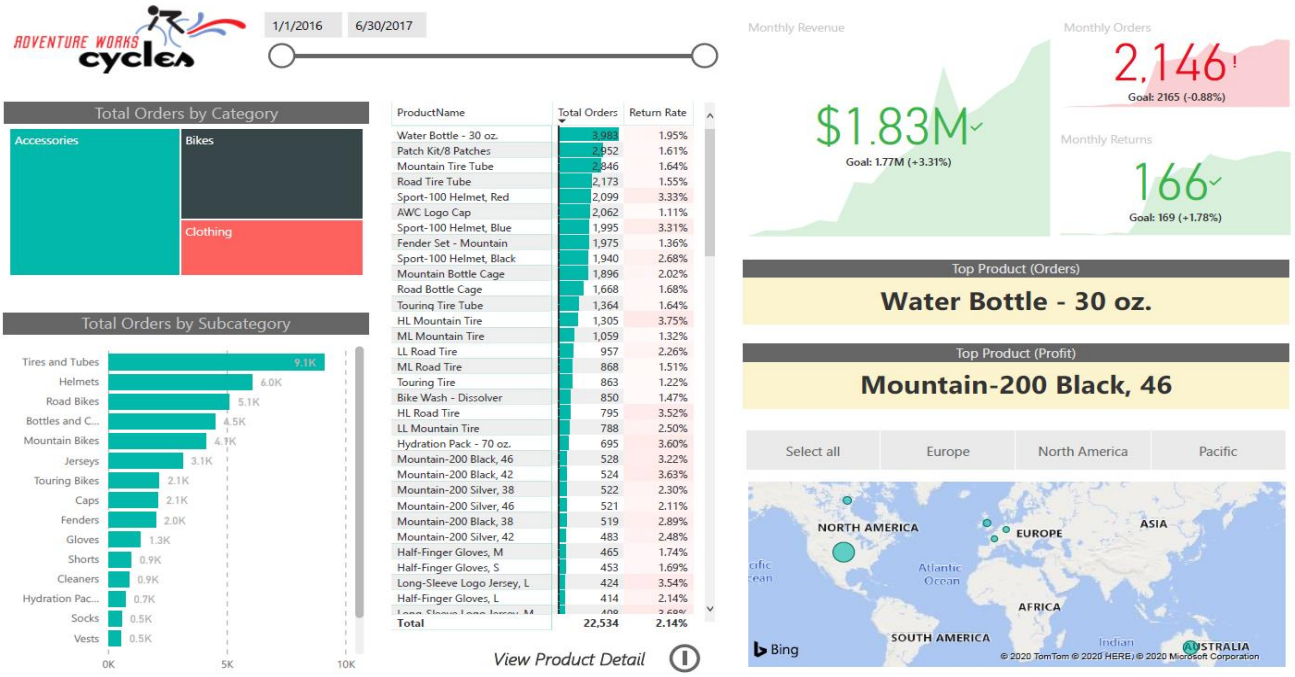
Data-Driven Insights

- What happened (Sold power was greater than the generated power in autumn months)
- Why did it happen (Billing cycle inaccuracies caused shifting months' bills)
- How to react (Utilization of smart meters to collect the readings at the end of the month in order to remove any human errors or delays in bills processing)



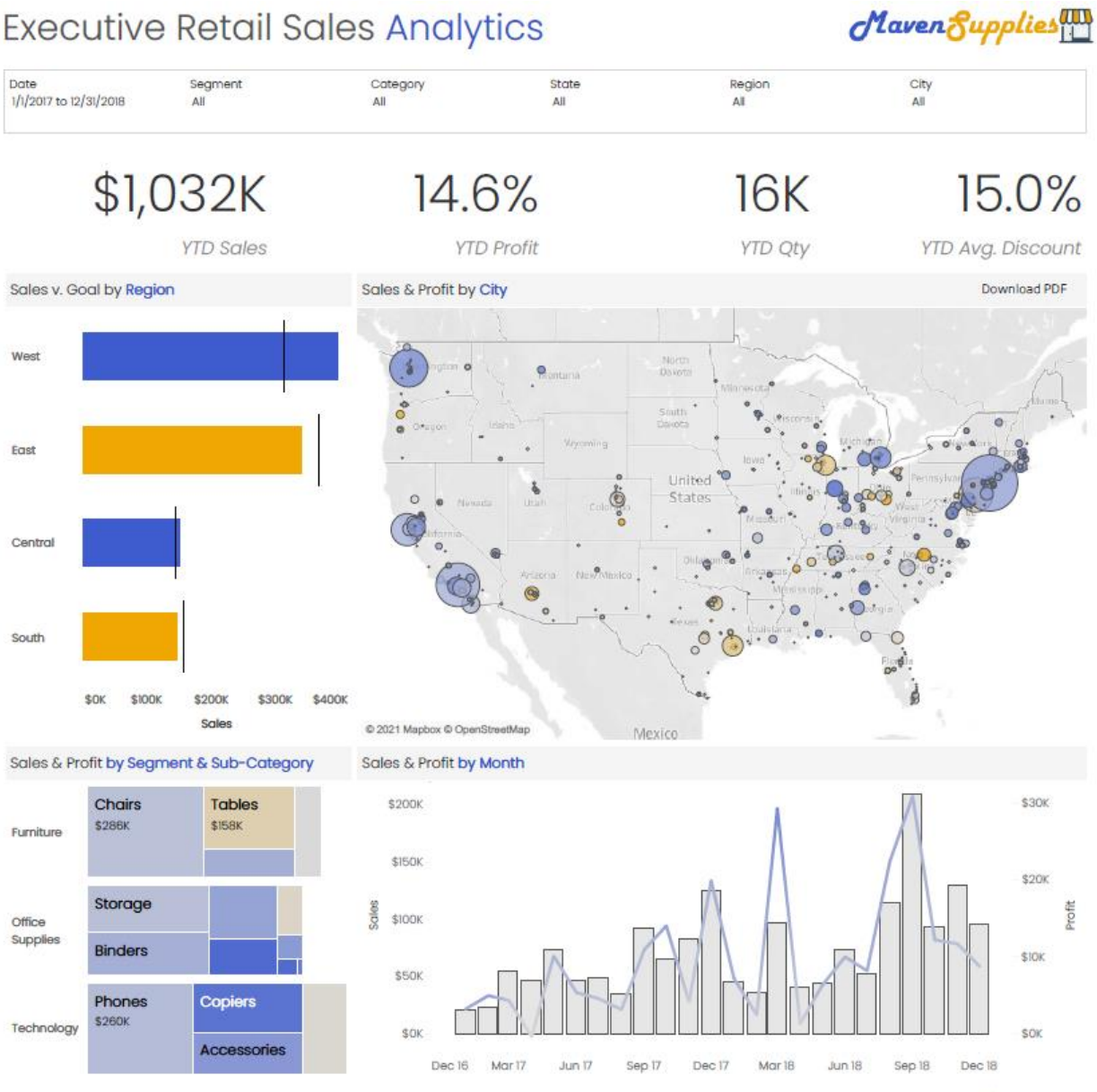
Project done using Microsoft Power BI

Adventure Works Cycles, a global manufacturing company, needed a way to track KPIs (sales, revenue, profit, returns), compare regional performance, analyze product level trends and forecasts, and identify high value customers.

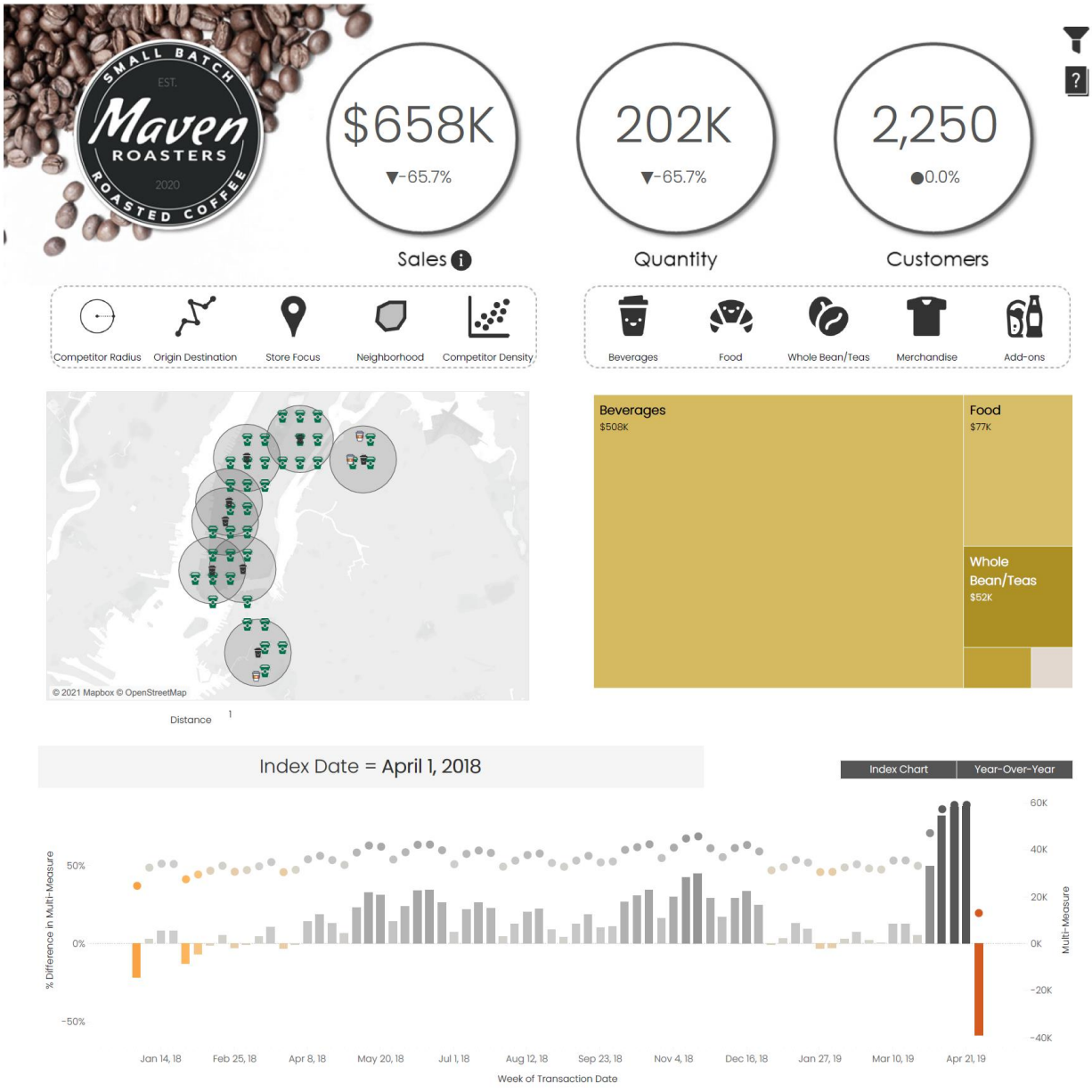


Projects done using Tableau

Maven Supplies, a cutting-edge office supplies company, needed a way to track KPIs (sales, profit, units, returns), compare performance across markets, analyze category profitability, and identify high value customers.

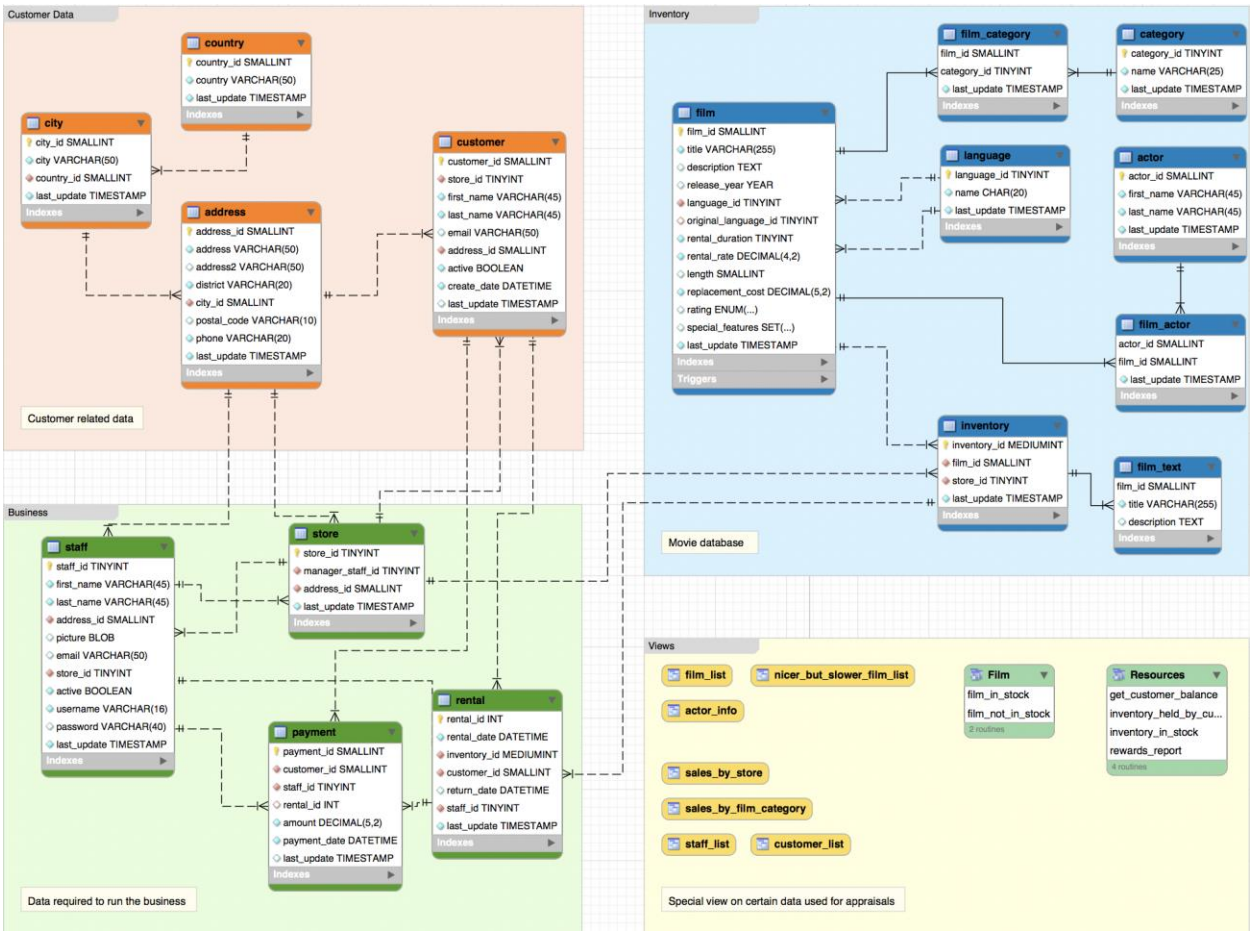


Maven Roasters is a coffee roasting company. They need a way to track KPIs (sales, profit, units, returns), compare performance across markets, analyze category profitability, and identify high value customers.



Projects done using MySQL

Maven Movies insurance policy is up for renewal and the insurance company's underwriters need some updated information before they will issue a new policy.

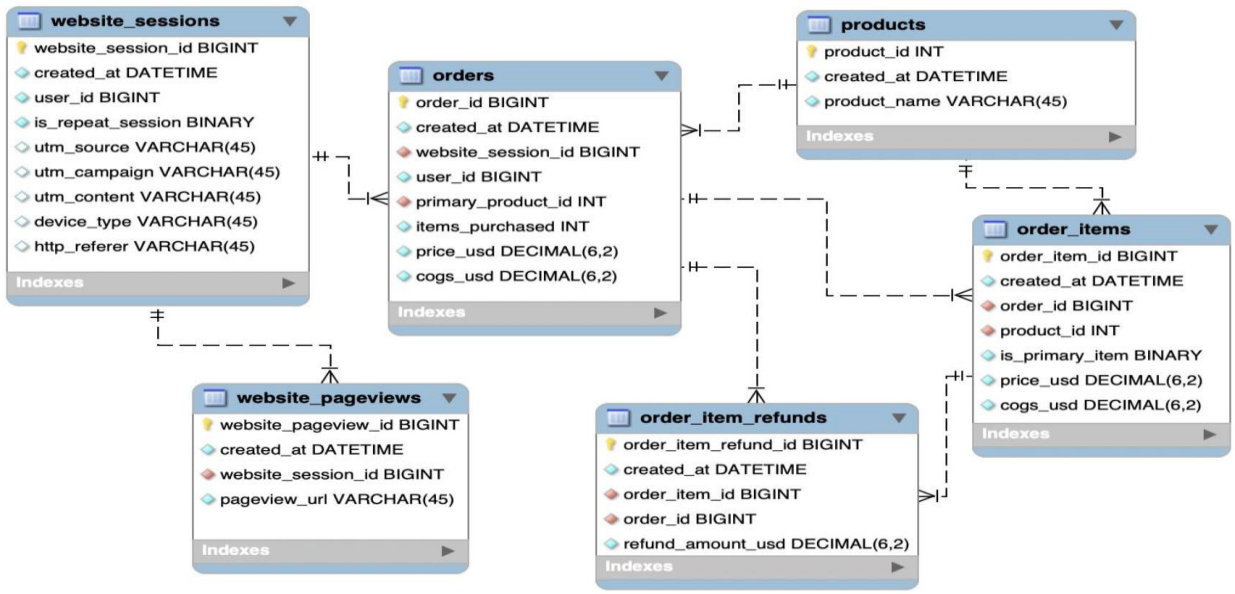


Questions Answered:

- A list of all staff members, including their first and last names, email addresses, and the store identification number where they work.
- What is the count of active customers for each of your stores?
- What is the count of inventory items held at each of your two stores?
- What is the count of all customer email addresses stored in the database?
- How diverse your film offering is as a means of understanding how likely you are to keep customers engaged in the future?
- What is the average payment you process, as well as the maximum payment you have processed?
- What is the replacement cost of your films?

- What your customer base looks like?
- A list of the managers' names at each store, with the full address of each property (street address, district, city, and country
- How many inventory items you have with each rating at each store?
- A list of each inventory item you have stocked.
- How diversified the inventory is in terms of replacement cost?
- What is the number of films, as well as the average replacement cost, and total replacement cost?
- A list of all customer names, which store they go to, whether or not they are currently active, and their full addresses street address, city, and country.
- A list of advisor and investor names in one table? note whether they are an investor or an advisor, and for the investors, it would be good to include which company they work with.
- How much your customers are spending with you, and also to know who your most valuable customers are. It would be great to see this ordered on total lifetime value?
- How well you have covered the most awarded actors?

Maven Fuzzy Factory is an eCommerce company. I was tasked to analyze and optimize marketing channels, measure and test website conversion performance, and use data to understand the impact of new product launches.



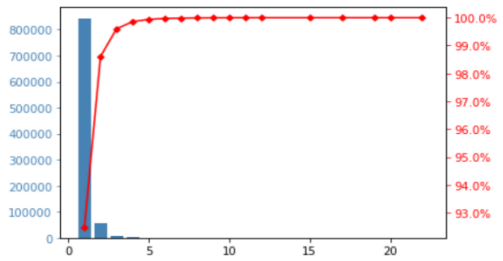
Questions Answered:

- Traffic source analysis
- Traffic conversion rate (CVR)
- Bid optimization analysis
- Traffic source trending
- Website performance analysis
- Landing page performance & testing
- Calculating bounce rates
- Conversion funnel analysis & testing
- Marketing channel portfolio optimization
- Direct traffic analysis
- Business patterns & seasonality analysis
- Product-level sales analysis
- Product launch sales analysis
- Product-level website analysis
- Cross-sell analysis
- Portfolio expansion analysis
- Product refund analysis
- Repeat behavior analysis

Projects done using Python

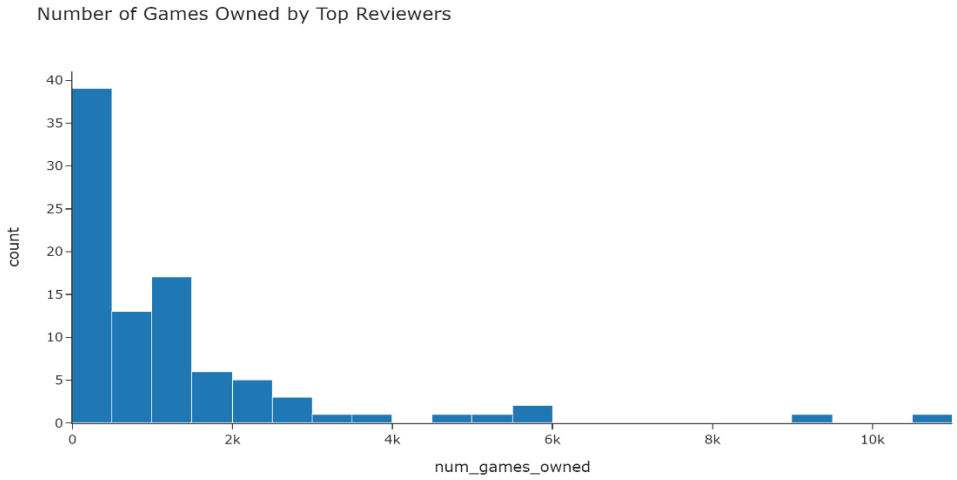
Top reviewers' analysis for Steam gaming platform for more than **1 million rows of data.**

In [23]:
```
# we will create a pareto chart to prove our point that the vast majority of reviewers have only 1 review
import matplotlib.pyplot as plt
from matplotlib.ticker import PercentFormatter

# define aesthetics for plot
color1= 'steelblue'
color2= 'red'
line_size = 4

# create a basic bar plot
fig, ax=plt.subplots()
ax.bar(df_reveiw_count['review_id'], df_reveiw_count['steam_id'], color=color1)


# add cumulative percentage line to plot
ax2=ax.twinx()
ax2.plot(df_reveiw_count['review_id'], df_reveiw_count['cumperc'], color=color2, marker="D", ms=line_size)
ax2.yaxis.set_major_formatter(PercentFormatter())

# specify axis colors
ax.tick_params(axis='y', colors=color1)
ax2.tick_params(axis='y', colors = color2)

# display pareto chart
plt.show()
```



Reviewer Analysis

In [49]:
```
# Number of Games Owned by Top Reviewers
number_games_owned=top_reviewers_df.groupby('steam_id')['num_games_owned'].max().reset_index()
px.histogram(number_games_owned, 'num_games_owned',nbins=50, template='simple_white', title='Number of Games Owned by Top Rev
```



Data cleaning & correlation analysis for **IMDB movies** dataset.

In [343]:
```
# Dropping the rows with null values in the following 'released', 'votes', 'writer', 'star' , 'company' columns
df2=df2.dropna(subset=['released', 'votes', 'writer', 'star', 'company'])
df2.head()
```

Out[343]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	company	runtime
0	The Shining	R	Drama	1980	June 13, 1980 (United States)	8.4	927000.0	Stanley Kubrick	Stephen King	Jack Nicholson	United Kingdom	19000000.0	46998772.0	Warner Bros.	146.0
1	The Blue Lagoon	R	Adventure	1980	July 2, 1980 (United States)	5.8	65000.0	Randal Kleiser	Henry De Vere Stacpoole	Brooke Shields	United States	4500000.0	58853106.0	Columbia Pictures	104.0
2	Star Wars: Episode V - The Empire Strikes Back	PG	Action	1980	June 20, 1980 (United States)	8.7	1200000.0	Irvin Kershner	Leigh Brackett	Mark Hamill	United States	18000000.0	538375067.0	Lucasfilm	124.0
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams	Jim Abrahams	Robert Hays	United States	3500000.0	83453539.0	Paramount Pictures	88.0
4	Caddyshack	R	Comedy	1980	July 25, 1980 (United States)	7.3	108000.0	Harold Ramis	Brian Doyle-Murray	Chevy Chase	United States	6000000.0	39846344.0	Orion Pictures	98.0

In [344]:
```
df2.isnull().sum()
```

Out[344]:
```
name           0
rating         0
genre          0
year           0
released       0
score          0
votes          0
director       0
writer         0
star           0
country        1
budget      2157
gross        180
company        0
runtime        2
dtype: int64
```

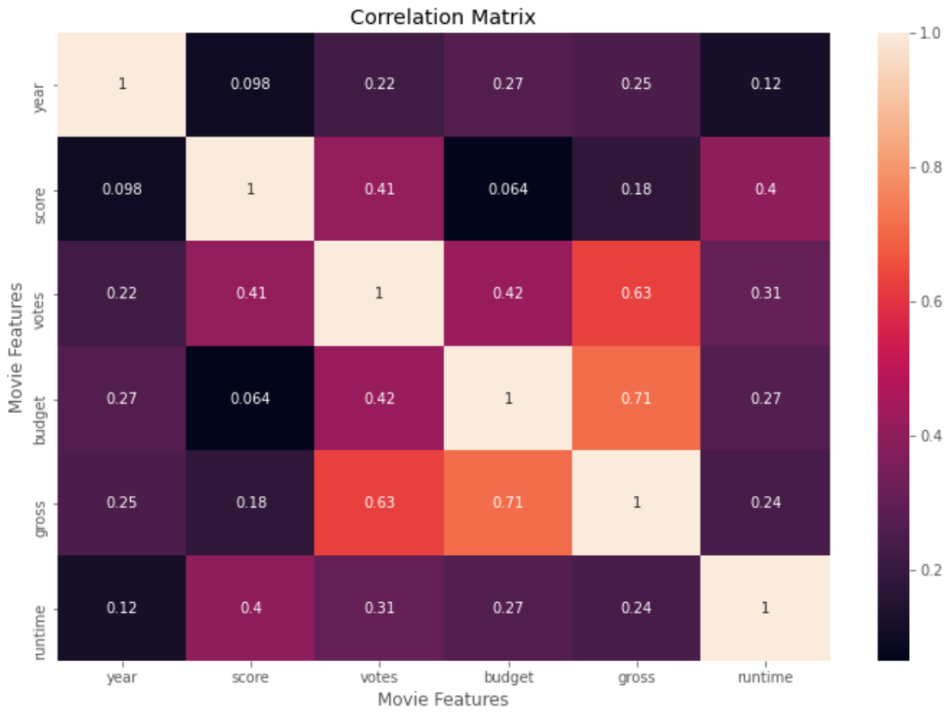
In [345]:
```
# Filling Null values in the "score" , "runtime", "gross", "budget"  columns with the mean value of their column
df2=df2.fillna({'score': df2['score'].mean(),
                'runtime': df2['runtime'].mean(),
                'gross': df2['gross'].mean(),
                'budget': df2['budget'].mean()})
df2.head()
```

In [360]:
```
# Creating a correlation matrix
df2.corr()
```

Out[360]:

	year	score	votes	budget	gross	runtime
year	1.000000	0.097791	0.222963	0.266582	0.252166	0.119357
score	0.097791	1.000000	0.409489	0.064297	0.182902	0.399951
votes	0.222963	0.409489	1.000000	0.420850	0.628740	0.308952
budget	0.266582	0.064297	0.420850	1.000000	0.711526	0.265199
gross	0.252166	0.182902	0.628740	0.711526	1.000000	0.241347
runtime	0.119357	0.399951	0.308952	0.265199	0.241347	1.000000

In [361]:
```
# Creating a heatmap from the correlation matrix

correlation_matrix=df2.corr(method='pearson')
sns.heatmap(correlation_matrix, annot=True)
plt.title('Correlation Matrix')
plt.xlabel('Movie Features')
plt.ylabel('Movie Features')
plt.show()
```



Credit risk model data preprocessing to estimate the probability of default.

Filling Sub Grade

In [233]:
```
# we can assign appropriate values for empty cells in the "sub-grade" column as it is related to the "Grade" column
# first condition check whether the value in "sub grade" column is missing,
# second condition indicates that in case the 1st condition is staisfied then,
# it checks whether the value in the "grade" column is equal to the value of the iterator variable for the given pass of the
# then is assigns the lowest/worst 5th sub-grade of that grade,
# for example, if there is missing value in the column "Sub-grade", then check its corresponding value in the "grade" column
# then assign (B5) to that missing value

for i in np.unique(loan_data_strings[:,3])[1:]: # loops goes through alll the unique grades after the first one (the empty sp
    loan_data_strings[:,4] = np.where((loan_data_strings[:,4] == '') & (loan_data_strings[:,3] == i),
                                      i + '5',
                                      loan_data_strings[:,4])
```

In [234]:
```
# There are still (9) rows in the column where we have neither the "sub-grade" nor the "grade"
# since there are 10,000 values in the columns, those 9 values can be dropped.
# However, since the applicant is withholding information, he/she should be penalized ( as the forms are filled manually by t
# Hence, we are going to creat a new sub-grade that is even lower than G5 for those applicants
np.unique(loan_data_strings[:,4], return_counts = True)
```

Out[234]:
```
(array(['', 'A1', 'A2', 'A3', 'A4', 'A5', 'B1', 'B2', 'B3', 'B4', 'B5', 'C1', 'C2', 'C3', 'C4',
        'C5', 'D1', 'D2', 'D3', 'D4', 'D5', 'E1', 'E2', 'E3', 'E4', 'E5', 'F1', 'F2', 'F3', 'F4',
        'F5', 'G1', 'G2', 'G3', 'G4', 'G5'], dtype='<U69'),
 array([  9, 285, 278, 239, 323, 592, 509, 517, 530, 553, 633, 629, 567, 586, 564, 577, 391, 267,
        250, 255, 288, 235, 162, 171, 139, 160,  94,  52,  34,  43,  24,  19,  10,   3,   7,   5],
       dtype=int64))
```

In [235]:
```
loan_data_strings[:,4] = np.where((loan_data_strings[:,4] == ''),
                                  "H1",
                                  loan_data_strings[:,4])
```

In [236]:
```
np.unique(loan_data_strings[:,4], return_counts = True)
```

Out[236]:
```
(array(['A1', 'A2', 'A3', 'A4', 'A5', 'B1', 'B2', 'B3', 'B4', 'B5', 'C1', 'C2', 'C3', 'C4', 'C5',
        'D1', 'D2', 'D3', 'D4', 'D5', 'E1', 'E2', 'E3', 'E4', 'E5', 'F1', 'F2', 'F3', 'F4', 'F5',
        'G1', 'G2', 'G3', 'G4', 'G5', 'H1'], dtype='<U69'),
```

Data preprocessing of raw **absenteeism data** into meaningful quantitative information.

Creating a Dummy Variable for the "Reason for Absence" Column

In [13]:
```
reasons_columns=pd.get_dummies(df['Reason for Absence'])
reasons_columns.head()
```

Out[13]:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	21	22	23	24	25	26	27	28
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0

In [14]:
```
# To check whether there is one reason only per each row, we create the "Checking" column by summing along the row
# 0 means there are missing values, 1 means there is only one reason, 2 means there are multiple reasons
reasons_columns['Checking']=reasons_columns.sum(axis=1)
reasons_columns.head()
```

Out[14]:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	21	22	23	24	25	26	27	28	Checking
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1

Projects done using Excel Pivot Table

U.S. Voter Demographics - 2012 population and voter registration data from the U.S. Census Bureau (sample=255)

- How many states had a Voter Population **%** below 55%? Which states?
- How many confirmed voters in California were over 65 years old in 2012? What percentage does that represent out of the total confirmed voters in California? What percentage out of the confirmed voters in the entire country?

- What percentage of the citizen population do 45 to 64-year-old represent? What percentage of the confirmed voter population?
- Which state had the highest voter turnout rate? What about among 18-24-year-old voters specifically?
- As a politician seeking to improve voter turnout rates among young adults (18-24), which particular states would you target first?

Salary information from San Francisco government employees, 2011-2013 (sample=24,285)

- Who were the 5 employees who earned the highest Base Pay in 2011?
- How many job titles earned *only* Other pay in 2012?
- Among employees with >=$100k Base Pay in 2012, Did any employee earn more than 50% of their salary from Other Pay? If so, who?

- How many employees held some sort of Curator position in either 2012 or 2013? Among those, who earned the highest average base pay?

Shark attack records from 1900-2016 (sample=5292)

- Which 3 countries had the highest number of reported attacks over the past 5 years (2012-2016)? During this period, what % of reported attacks occurred in Spain?
- What are the top 5 areas by count of case number and by country? Where in South Africa were shark attacks most frequently reported over these 5 years?
- What % of attacks in New Zealand were unprovoked? How many cases?

A 3-month sample of stock market data for 500 publicly traded companies (sample=29,440)

- On which date in the sample did Amazon (AMZN) see the largest price spread?
- When did Google (GOOG) see the largest day-over-day gain in trading volume? The largest drop?

Major League Baseball team statistics by season, 1995-2015 (sample=624)

- Which team had the highest Net Run total over the entire sample? What about just the 2015 season?
- Which years did the Red Sox win the Division? The Wild Card? The World Series?
- In which season did overall home run totals decrease the most Y-o-Y?

Burrito ratings and Yelp reviews from 65 San Diego restaurants in 2016 (sample=237)

- Compare average ratings for Tortilla, Temp, Fillings, Synergy, and Wrap Quality, by Location
- How many locations recorded >2 ratings?
- Which location has the lowest and highest total average score?

Daily weather conditions in Boston, Massachusetts from Jan Dec 2016 (sample=363)

- How many days in 2016 were categorized as Clear vs. Rain vs. Snow?
- What was the average temperature on clear days vs. snowy days? What about the average max temperature?
- What percent of September days are clear?
- How often did it snow in January 2016, as a percentage of the month?
- In how many months of 2016 did it not snow at all?

Spartan Race Facebook posts from Aug-Oct 2016 (sample=393)

- On which date did Spartan Race post most often? What types of posts were they?
- Which date showed the highest total engagement volume? Highest engagements per post? On the date with the highest engagements per post, which one drove the strongest response?
- Which post types tend to be most "sharable"? What time of day do people tend to share most?
- What time of day are users LEAST likely to actively engage with Spartan Race Facebook posts? How many active engagements per post during that time of day?

Details and descriptions for 7,000+ mobile apps available on the Apple App Store (7,197)

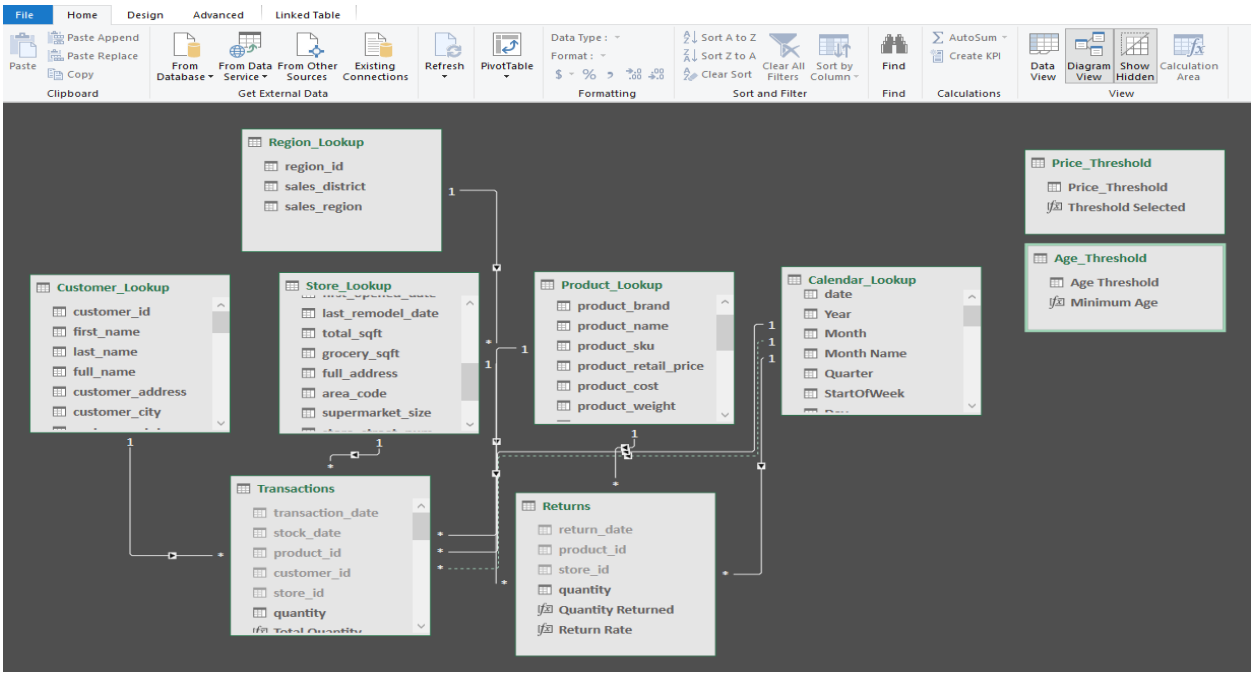
- How many apps are categorized as high-volume?
- What percentage of all apps are high-volume, and what percentage of all ratings do those high-volume apps account for?
- How do average ratings compare between high-volume and low-volume apps?
- What are the Top 10 apps by total ratings?
- Among high-volume Entertainment apps specifically, which app saw the largest rating improvement with the current version?

Wine Tasting Scores (sample=129,971)

- Which sub region produced the highest-rated & highest-priced wines, on average?
- Which sub region has the most tastings recorded in the dataset?
- On average, which country has generated the highest & lowest average ratings?
- How do the ratings compare across countries for Pinot Grigio varieties specifically?
- What's the cheapest Italian wine rated over 95 points? Who reviewed that particular wine, and how did he/she describe it?

Project done using Excel Power Query

Supermarket chain “Food Mart”, In addition to daily transactional records from 1997-1998, the data set included information about products, customers, stores, and regions.

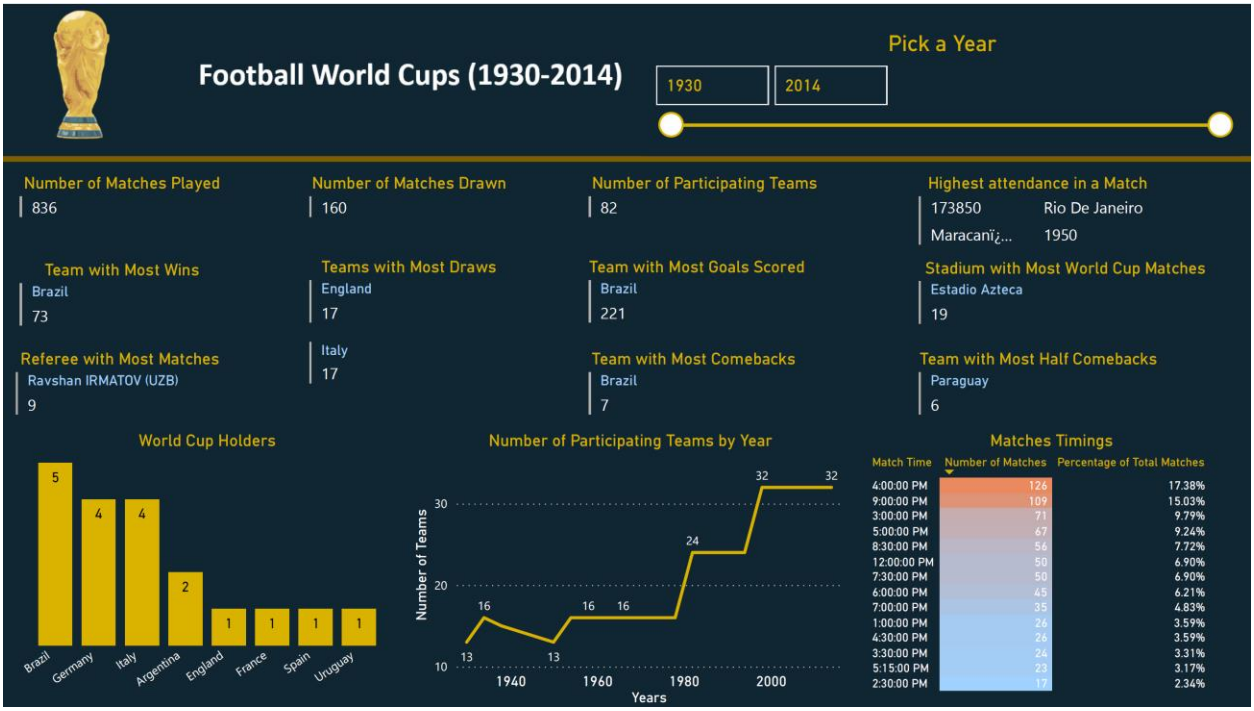


Additional Projects

Football World Cup:

Data from all World Cup results (1930-2014), details for each match (Date, Stage, City, Ref, Home Team, Away Team, Goals, Goals at Half), and details for events (goals, assists) by athlete.

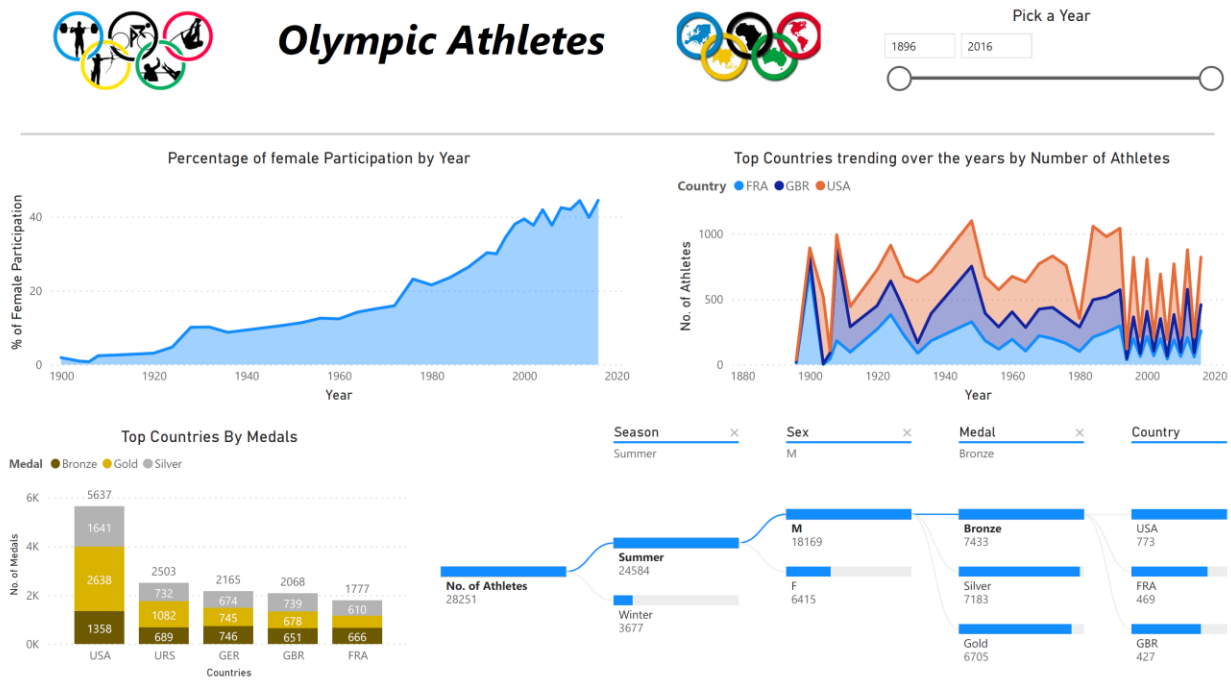
Number of fields: 39, Number of records: 3656



Olympic Games:

Historical data on the modern Olympic Games, from Athens 1896 to Rio 2016. Each row corresponds to an individual athlete competing in an individual event, including the athlete's name, sex, age, height, weight, country, and medal, and the event's name, sport, games, year, and city.

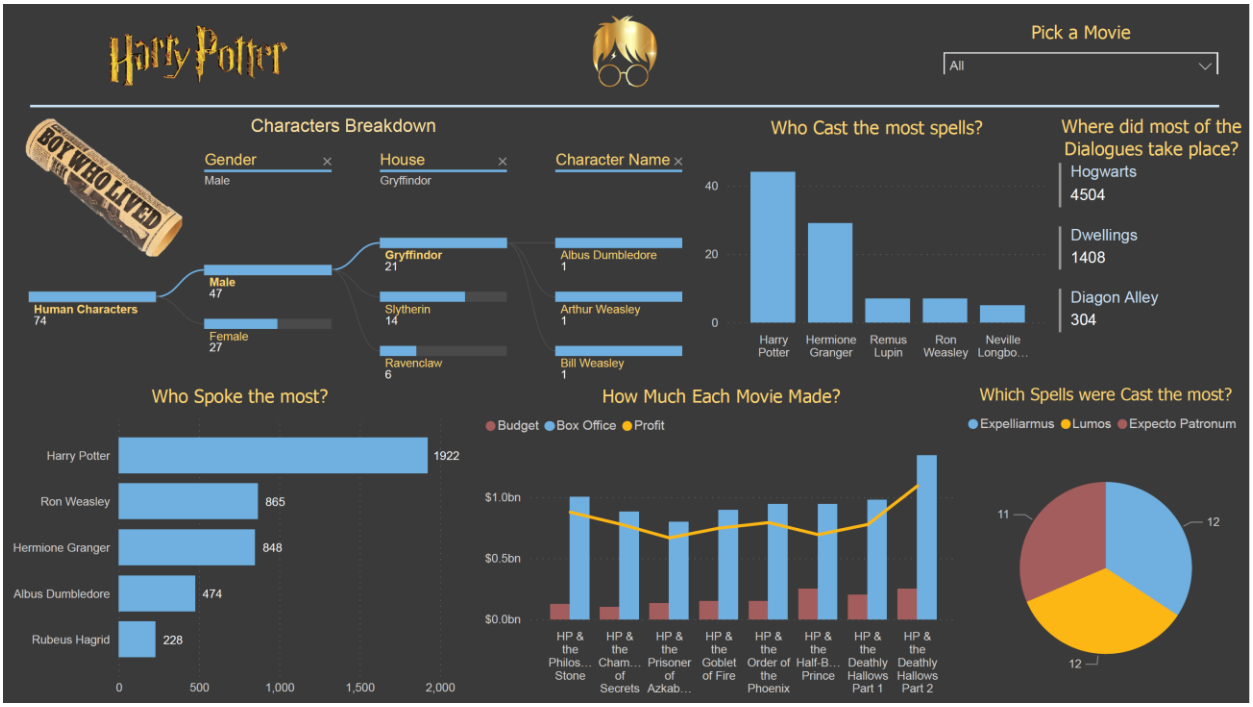
Number of fields= 16, Number of records= 271,116



Harry Potter Movies:

Character dialogues for all 8 movies in the Harry Potter franchise, including additional information about each movie, its chapters, characters, places, and spells.

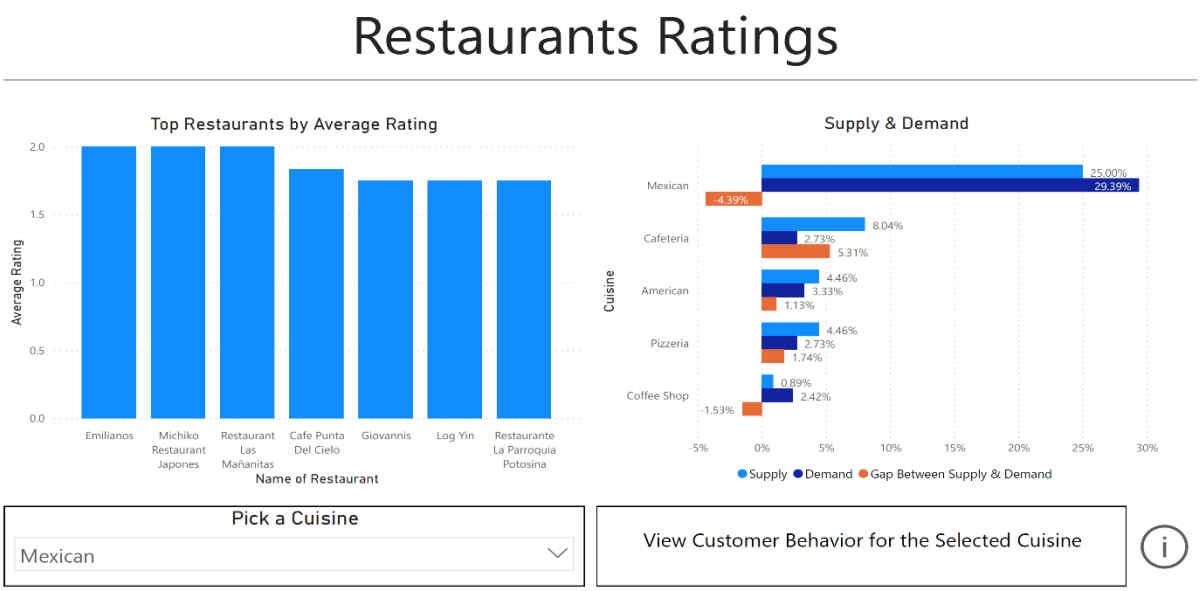
Number of fields: 27, Number of records: 7987

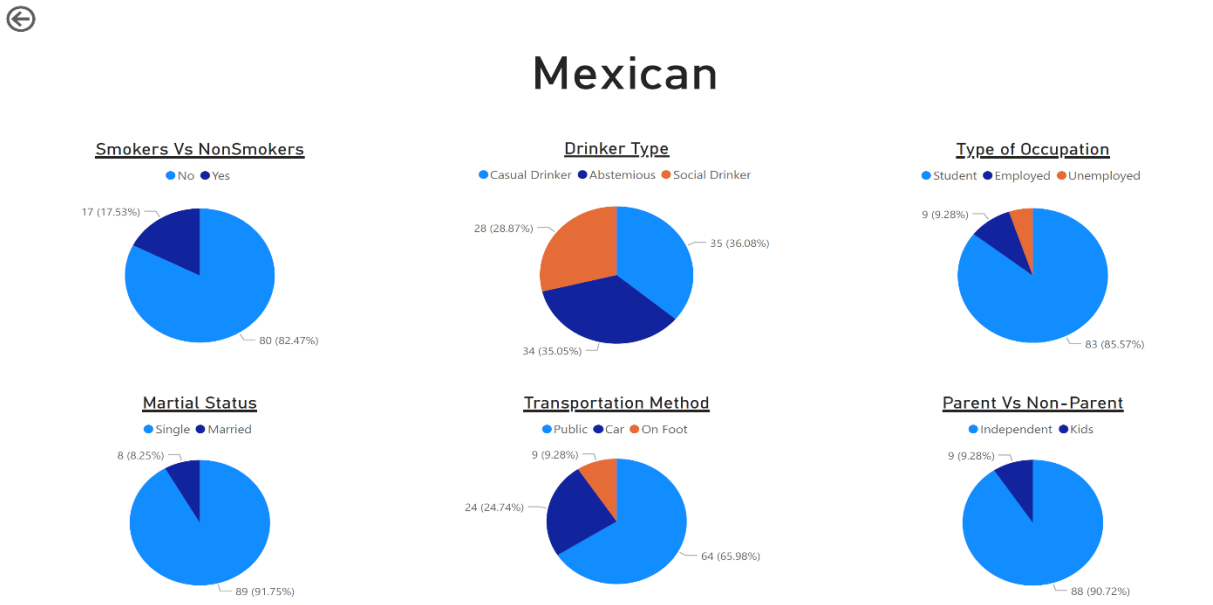


Restaurants Ratings:

Restaurant ratings in Mexico by real consumers from 2012, including additional information about each restaurant and their cuisines, and each consumer and their preferences.

Number of fields: 24, Number of records: 1653

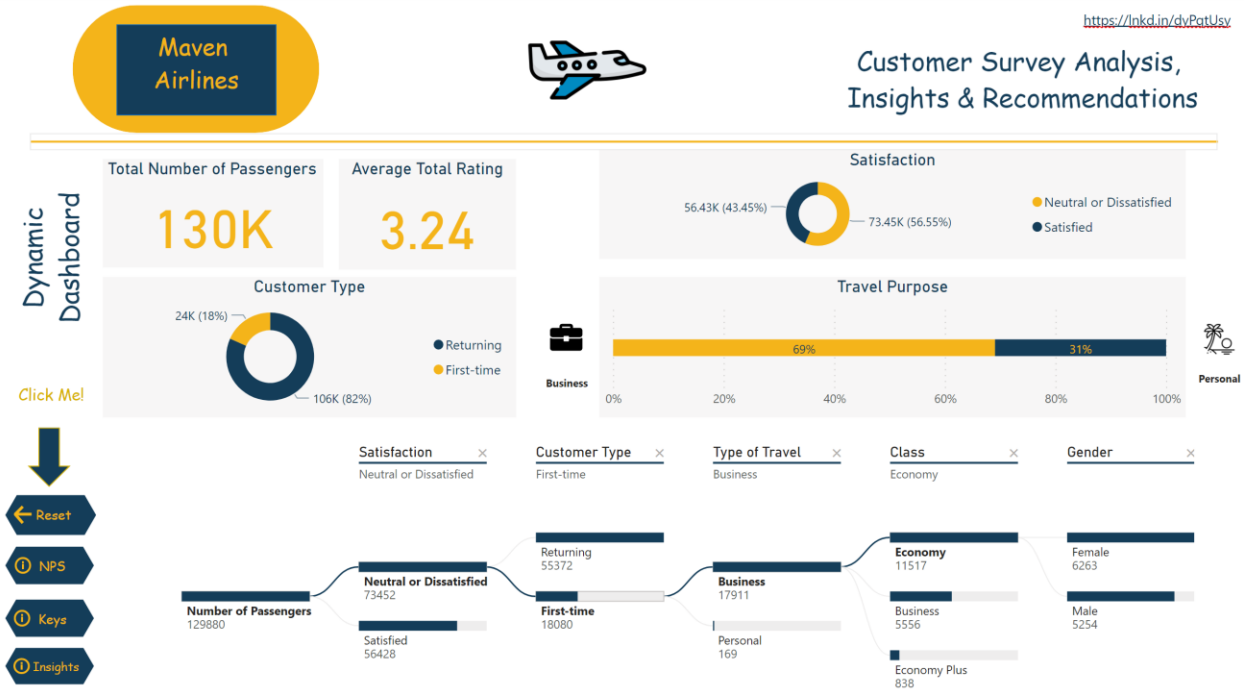




Maven Airlines:

Customer satisfaction scores from 120,000+ airline passengers, including additional information about each passenger, their flight, and type of travel, as well as their evaluation of different factors like cleanliness, comfort, service, and overall experience.

Number of fields: 24, Number of records: 129880



Telecom Customer Churn:

Churn data for a fictional Telecommunications company that provides phone and internet services to 7,043 customers in California, and includes details about customer demographics, location, services, and current status. The main task is to improve retention by identifying high value customers and churn risks.

Number of fields: 39, Number of records: 7043