

# Summary of Case Studies towards Master's of Computer Science

## Specialization in Machine Learning & Artificial Intelligence

Suman Debnath

### 1. Business Case: Aerofit - Descriptive Statistics & Probability

#### Project Details

 Aerofit - Descriptive Statistics & Probability

#### Learning Summary

- The Aerofit project utilized descriptive analytics to explore customer preferences for treadmill models.
- Employed Python libraries: pandas, numpy, matplotlib, seaborn, and scipy for customer data analysis.
- Analyzed customer profiles, usage patterns, and income levels for product recommendation optimization.
- Emphasized the importance of data-driven decision-making and extracting actionable insights.
- Highlighted the practical application of statistical analysis in business, despite not mentioning a specific algorithm.

#### Introduction

Aerofit is a leading brand in the field of fitness equipment. Aerofit provides a product range including machines such as treadmills, exercise bikes, gym equipment, and fitness accessories to cater to the needs of all categories of people.

#### Business Problem

The market research team at AeroFit wants to identify the characteristics of the target audience for each type of treadmill offered by the company, to provide a better recommendation of the treadmills to the new customers. The team decides to investigate whether there are differences across the product with respect to customer characteristics.

1. Perform descriptive analytics to **create a customer profile** for each AeroFit treadmill product by developing appropriate tables and charts.
2. For each AeroFit treadmill product, construct **two-way contingency tables** and compute all **conditional and marginal probabilities** along with their insights/impact on the business.

### Dataset

Link: [Dataset\\_link](#)

The dataset have the following fields:

## Customer based profiling

### KP281

- Easily affordable entry level product, which is also the maximum selling product.
- KP281 is the most popular product among the entry level customers.
- This product is easily afforded by both Male and Female customers.
- Average distance covered in this model is around 70 to 90 miles.
- Product is used 3 to 4 times a week.
- Most of the customer who have purchased the product have rated Average shape as the fitness rating.
- Younger to Elder beginner level customers prefer this product.
- Single female & Partnered male customers bought this product more than single male customers.
- Income range between 39K to 53K have preferred this product.

### KP481

- This is an Intermediate level Product.

- KP481 is the second most popular product among the customers.
- Fitness Level of this product users varies from Bad to Average Shape depending on their usage.
- Customers Prefer this product mostly to cover more miles than fitness.
- Average distance covered in this product is from 70 to 130 miles per week.
- More Female customers prefer this product than males.
- Probability of Female customer buying KP481 is significantly higher than male.
- KP481 product is specifically recommended for Female customers who are intermediate user.
- Three different age groups prefer this product - Teen, Adult and middle aged.
- Average Income of the customer who buys KP481 is 49K.
- Average Usage of this product is 3 days per week.
- More Partnered customers prefer this product.
- There are slightly more male buyers of the KP481.
- The distance travelled on the KP481 treadmill is roughly between 75 - 100 Miles. It is also the 2nd most distance travelled model.
- The buyers of KP481 in Single & Partnered, Male & Female are same.
- The age range of KP481 treadmill customers is roughly between 24-34 years.

### **KP781**

- Due to the High Price & being the advanced type, customer prefers less of this product.
- Customers use this product mainly to cover more distance.
- Customers who use this product have rated excelled shape as fitness rating.
- Customer walk/run average 120 to 200 or more miles per week on his product.
- Customers use 4 to 5 times a week at least.
- Female Customers who are running average 180 miles (extensive exercise) , are using product KP781, which is higher than Male average using same product.
- Probability of Male customer buying Product KP781(31.73%) is way more than female(9.21%).

- Probability of a single person buying KP781 is higher than Married customers. So , KP781 is also recommended for people who are single and exercises more.
- Middle aged to higher age customers tend to use this model to cover more distance.
- Average Income of KP781 buyers are over 75K per annum
- Partnered Female bought KP781 treadmill compared to Partnered Male.
- Customers who have more experience with previous aerofit products tend to buy this product
- This product is preferred by the customer where the correlation between Education and Income is High.

## Recommendation

- Female who prefer exercising equipments are very low here. Hence, we should run a marketing campaign on to encourage women to exercise more
- KP281 & KP481 treadmills are preferred by the customers whose annual income lies in the range of 39K - 53K Dollars. These models should promoted as budget treadmills.
- As KP781 provides more features and functionalities, the treadmill should be marketed for professionals and athletes.
- KP781 product should be promoted using influencers and other international athletes.
- Research required for expanding market beyond 50 years of age considering health pros and cons.
- Provide customer support and recommend users to upgrade from lower versions to next level versions after consistent usages.
- KP781 can be recommended for Female customers who exercises extensively along with easy usage guidance since this type is advanced.
- Target the Age group above 40 years to recommend Product KP781.
- Education with 14 to 16 years have tendency to buy more of KP281 models

## Conclusion

- KP 281 model is the most purchased model (44.4%) then KP 481 (33.3%). KP 781 is the least sold model (22.2%).

- There are more Male customers (57.8%) than Female customers (42.2%).
- Average Usage of Males is more than Average usage of Females.
- Customers buying treadmill are younger and average age of customer is 28.
- Most of the customers earns less than 70K and prefer KP 281 & KP 481 models.
- 59.4% of the customers who purchased treadmill are partnered.
- Customers average education is 16.

## Future

- For KP281: Concentrate advertising broadly across gender and marital status towards individuals with annual income less than \$75,000, with some college education or a bachelor's degree, who are unfit or average fitness and in their 20s or 30s.
- For KP481 : Concentrate advertising broadly across gender and marital status towards individuals with annual income less than \$75,000, with some college education or a bachelor's degree, who are unfit or average fitness and in their 20s or 30s.
- For KP781: Concentrate advertising towards males who are average fitness to very fit, have a bachelors degree or advanced education, and are in their 20s or 30s.
- There may be untapped potential for targeting customers in the 40s and beyond age group, which appear to be an underserved population. Analysis indicates more than just outlying purchases of KP 781 .
- Individuals with only a high school education also appear to be an underserved population. Likely best candidates for KP 281 or KP 481 due to annual income constraints.

## 2. Yulu - Hypothesis Testing

### Project Details

 [Yulu - Hypothesis Testing](#)

# Learning Summary

- The Yulu project analyzed demand for shared electric cycles in India through hypothesis testing and descriptive analytics.
- It evaluated the impact of seasonality, holidays, working days, and weather conditions on cycle rental counts.
- Insights showed significant patterns in rental demand related to seasons and weather, guiding marketing and inventory management.
- Showcasing hypothesis testing's role in understanding consumer behavior and enhancing service offerings.

## Introduction

Yulu is India's leading micro-mobility service provider, which offers unique vehicles for the daily commute. Starting off as a mission to eliminate traffic congestion in India, Yulu provides the safest commute solution through a user-friendly mobile app to enable shared, solo and sustainable commuting.

Yulu zones are located at all the appropriate locations (including metro stations, bus stands, office spaces, residential areas, corporate offices, etc) to make those first and last miles smooth, affordable, and convenient!

Yulu has recently suffered considerable dips in its revenues. They have contracted a consulting company to understand the factors on which the demand for these shared electric cycles depends. Specifically, they want to understand the factors affecting the demand for these shared electric cycles in the Indian market.

### Business Problem

The company wants to know:

- Which variables are significant in predicting the demand for shared electric cycles in the Indian market?
- How well those variables describe the electric cycle demands

### Dataset

Dataset link: [yulu\\_data.csv](#)

The dataset have the following fields:

`datetime` : datetime

`season` : season (1: spring, 2: summer, 3: fall, 4: winter)

**holiday** : whether day is a holiday or not (extracted from <http://dchr.dc.gov/page/holiday-schedule>)

**workingday** : if day is neither weekend nor holiday is 1, otherwise is 0.

**weather** : - Clear, Few clouds, partly cloudy, partly cloudy - Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist - Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds - Heavy Rain + Ice Pellets + Thunderstorm + Mist, Snow + Fog

**temp** : temperature in Celsius

**atemp** : feeling temperature in Celsius

**humidity** : humidity

**windspeed** : wind speed

**casual** : count of casual users

**registered** : count of registered users

**count** : count of total rental bikes including both casual and registered

## Summary

- The data is given from Timestamp('2011-01-01 00:00:00') to Timestamp('2012-12-19 23:00:00'). The total time period for which the data is given is '718 days 23:00:00'.
- Out of every 100 users, around 19 are casual users and 81 are registered users.
- The mean total hourly count of rental bikes is 144 for the year 2011 and 239 for the year 2012. An annual growth rate of 65.41 % can be seen in the demand of electric vehicles on an hourly basis.
- There is a seasonal pattern in the count of rental bikes, with higher demand during the spring and summer months, a slight decline in the fall, and a further decrease in the winter months.
- The average hourly count of rental bikes is the lowest in the month of January followed by February and March.
- There is a distinct fluctuation in count throughout the day, with low counts during early morning hours, a sudden increase in the morning, a peak count in the afternoon, and a gradual decline in the evening and nighttime.
- More than 80 % of the time, the temperature is less than 28 degrees celcius.
- More than 80 % of the time, the humidity value is greater than 40. Thus for most of the time, humidity level varies from optimum to too moist.
- More than 85 % of the total, windspeed data has a value of less than 20.

- The hourly count of total rental bikes is the highest in the clear and cloudy weather, followed by the misty weather and rainy weather. There are very few records for extreme weather conditions.
- The mean hourly count of the total rental bikes is statistically similar for both working and non- working days.
- There is statistically significant dependency of weather and season based on the hourly total number of bikes rented.
- The hourly total number of rental bikes is statistically different for different weathers.
- There is no statistically significant dependency of weather 1, 2, 3 on season based on the average hourly total number of bikes rented.
- The hourly total number of rental bikes is statistically different for different seasons.

## Recommendation


- **Seasonal Marketing:** Since there is a clear seasonal pattern in the count of rental bikes, Yulu can adjust its marketing strategies accordingly. Focus on promoting bike rentals during the spring and summer months when there is higher demand. Offer seasonal discounts or special packages to attract more customers during these periods.
- **Time-based Pricing:** Take advantage of the hourly fluctuation in bike rental counts throughout the day. Consider implementing time-based pricing where rental rates are lower during off-peak hours and higher during peak hours. This can encourage customers to rent bikes during less busy times, balancing out the demand and optimizing the resources.
- **Weather-based Promotions:** Recognize the impact of weather on bike rentals. Create weather-based promotions that target customers during clear and cloudy weather, as these conditions show the highest rental counts. Yulu can offer weather-specific discounts to attract more customers during these favorable weather conditions.
- **User Segmentation:** Given that around 81% of users are registered, and the remaining 19% are casual, Yulu can tailor its marketing and communication strategies accordingly. Provide loyalty programs, exclusive offers, or personalized recommendations for registered users to encourage repeat business. For casual users, focus on providing a seamless rental experience and promoting the benefits of bike rentals for occasional use.
- **Optimize Inventory:** Analyze the demand patterns during different months and adjust the inventory accordingly. During months with lower rental counts such as January, February, and March, Yulu can optimize its inventory levels to avoid excess



bikes. On the other hand, during peak months, ensure having sufficient bikes available to meet the higher demand.

- **Improve Weather Data Collection:** Given the lack of records for extreme weather conditions, consider improving the data collection process for such scenarios. Having more data on extreme weather conditions can help to understand customer behavior and adjust the operations accordingly, such as offering specialized bike models for different weather conditions or implementing safety measures during extreme weather.
- **Customer Comfort:** Since humidity levels are generally high and temperature is often below 28 degrees Celsius, consider providing amenities like umbrellas, rain jackets, or water bottles to enhance the comfort and convenience of the customers. These small touches can contribute to a positive customer experience and encourage repeat business.
- **Collaborations with Weather Services:** Consider collaborating with weather services to provide real-time weather updates and forecasts to potential customers. Incorporate weather information into your marketing campaigns or rental app to showcase the ideal biking conditions and attract users who prefer certain weather conditions.
- **Seasonal Bike Maintenance:** Allocate resources for seasonal bike maintenance. Before the peak seasons, conduct thorough maintenance checks on the bike fleet to ensure they are in top condition. Regularly inspect and service bikes throughout the year to prevent breakdowns and maximize customer satisfaction.
- **Customer Feedback and Reviews:** Encourage customers to provide feedback and reviews on their biking experience. Collecting feedback can help identify areas for improvement, understand customer preferences, and tailor the services to better meet customer expectations.
- **Social Media Marketing:** Leverage social media platforms to promote the electric bike rental services. Share captivating visuals of biking experiences in different weather conditions, highlight customer testimonials, and engage with potential customers through interactive posts and contests. Utilize targeted advertising campaigns to reach specific customer segments and drive more bookings.
- **Special Occasion Discounts:** Since Yulu focusses on providing a sustainable solution for vehicular pollution, it should give special discounts on the occasions like Zero Emissions Day (21st September), Earth day (22nd April), World Environment Day (5th June) etc in order to attract new users.

## 3. OLA - Ensemble Learning

 No description has been provided for this image

### Project Details

 OLA - Ensemble Learning

### Learning Summary

- The OLA Ensemble Learning project aimed at predicting driver churn by analyzing demographics, tenure, and performance.
- Utilized algorithms: Random Forest, Bagging Classifier with Decision Trees, XGBoost Classifier, and Gradient Boosting Classifier. Gradient Boosting was particularly effective, with an accuracy of 0.8910 and ROC-AUC of 0.9448.
- Highlighted factors affecting driver retention, including joining year and total business value.
- Emphasized the importance of ensemble learning techniques for enhancing predictive accuracy in churn modeling.

### Introduction

- Recruiting and retaining drivers is seen by industry watchers as a tough battle for Ola.
- Churn among drivers is high and it's very easy for drivers to stop working for the service on the fly or jump to Uber depending on the rates.
- As the companies get bigger, the high churn could become a bigger problem. To find new drivers, Ola is casting a wide net, including people who don't have cars for jobs. But this acquisition is really costly.
- Losing drivers frequently impacts the morale of the organization and acquiring new drivers is more expensive than retaining existing ones.
- You are working as a data scientist with the Analytics Department of Ola, focused on driver team attrition.
- You are provided with the monthly information for a segment of drivers for 2019 and 2020 and tasked to predict whether a driver will be leaving the company or not based on their attributes like

- Demographics (city, age, gender etc.)
- Tenure information (joining date, Last Date)
- Historical data regarding the performance of the driver (Quarterly rating, Monthly business acquired, grade, Income)

## Column Profiling:

- MMMM-YY : Reporting Date (Monthly)
- Driver\_ID : Unique id for drivers
- Age : Age of the driver
- Gender : Gender of the driver – Male : 0, Female: 1
- City : City Code of the driver
- Education\_Level : Education level – 0 for 10+ ,1 for 12+ ,2 for graduate
- Income : Monthly average Income of the driver
- Date Of Joining : Joining date for the driver
- LastWorkingDate : Last date of working for the driver
- Joining Designation : Designation of the driver at the time of joining
- Grade : Grade of the driver at the time of reporting
- Total Business Value : The total business value acquired by the driver in a month (negative business indicates -cancellation/refund or car EMI adjustments)
- Quarterly Rating : Quarterly rating of the driver: 1,2,3,4,5 (higher is better)

## Summary

### Data Distribution:

- **Gender:**
  - Male: 1380
  - Female: 956
- **Churn Distribution:**
  - 1 (Churned): 1616 (67.87%)
  - 0 (Not Churned): 765 (32.13%)

### Random Forest:

- Train and test score: (0.8697, 0.8679)
- Highest feature importance: Joining year, followed by the number of records available in data, and total business value.
- Recall: 0.866

- Precision: 0.928
- F1-Score: 0.89

## Grid Search CV on Random Forest:

- Best parameters: ccp\_alpha=0.001, max\_depth=10, max\_features=7, n\_estimators=300
- Best score: 0.8881

## Bagging Classifier with Decision Trees:

- 50 Decision Trees, max\_depth=7, class\_weight="balanced"
- F1 Score: 0.9064
- Precision: 0.9388
- Recall Score: 0.8762
- Accuracy: 0.880

## XGBoost Classifier (Grid Search CV):

- Parameters: 'max\_depth': 2, 'n\_estimators': 100
- Test Scores:
  - Accuracy: 0.87
  - F1 Score: 0.90
  - Recall: 0.923
  - Precision: 0.884
- Highest feature importance: Joining year, followed by the number of records available in data, and total business value.

## Gradient Boosting Classifier (GBC):


- Train Score: 0.9144
- Test Score: 0.8910
- Accuracy Score: 0.8910
- ROC-AUC Score (test dataset): 0.9448
- Precision Score (test dataset): 0.9288
- Recall Score (test dataset): 0.9119
- F1 Score (test dataset): 0.9202

## Observations

- The probability of churn is higher in cases where the education level is 0 and 1, compared to 2.
- For drivers with a joining designation of 1, the probability of churn is higher.

- When the quarterly rating is 1, the probability of churn is significantly higher.
- A similar pattern is observed for drivers whose quarterly rating has increased throughout their tenure.
- Drivers who joined in 2018 and 2019 have a very high probability of churn compared to those who joined in 2020 or before 2018.

## 4. Delhivery - Feature Engineering

 No description has been provided for this image

### Project Details

 Delhivery - Feature Engineering

### Learning Summary

- The Delhivery project focused on optimizing logistics and delivery efficiency via data analysis.
- Concentrated on feature engineering, exploratory data analysis, and hypothesis testing to enhance understanding of delivery routes, times, and efficiency.
- Unveiled insights on delivery patterns, suggesting the use of small vehicles for intra-city and heavy trucks for long-distance deliveries.
- Analyzed 14,817 trips, applying statistical methods like two-sample t-tests to compare actual delivery times against planned, enhancing service delivery and operational efficiency.

### Introduction

Delhivery is the largest and fastest-growing fully integrated player in India by revenue in Fiscal 2021. They aim to build the operating system for commerce, through a combination of world-class infrastructure, logistics operations of the highest quality, and cutting-edge engineering and technology capabilities.

The Data team builds intelligence and capabilities using this data that helps them to widen the gap between the quality, efficiency, and profitability of their business versus their competitors.

#### Business Problem

The company wants to understand and process the data coming out of data engineering pipelines:

- Clean, sanitize and manipulate data to get useful features out of raw fields
- Make sense out of the raw data and help the data science team to build forecasting models on it

## Dataset

Dataset link: [delhivery\\_data.csv](#)

The dataset have the following fields:

- data - tells whether the data is testing or training data
- trip\_creation\_time - Timestamp of trip creation
- route\_schedule\_uuid - Unique Id for a particular route schedule
- route\_type - Transportation type
- FTL - Full Truck Load: FTL shipments get to the destination sooner, as the truck is making no other pickups or drop-offs along the way
- Carting: Handling system consisting of small vehicles (carts)
- trip\_uuid - Unique ID given to a particular trip (A trip may include different source and destination centers)
- source\_center - Source ID of trip origin
- source\_name - Source Name of trip origin
- destination\_cente - Destination ID
- destination\_name - Destination Name
- od\_start\_time - Trip start time
- od\_end\_time - Trip end time
- start\_scan\_to\_end\_scan - Time taken to deliver from source to destination
- is\_cutoff - Unknown field
- cutoff\_factor - Unknown field
- cutoff\_timestamp - Unknown field
- actual\_distance\_to\_destination - Distance in Kms between source and destination warehouse
- actual\_time - Actual time taken to complete the delivery (Cumulative)
- osrm\_time - An open-source routing engine time calculator which computes the shortest path between points in a given map (Includes usual traffic, distance through major and minor roads) and gives the time (Cumulative)
- osrm\_distance - An open-source routing engine which computes the shortest path between points in a given map (Includes usual traffic, distance through major and minor roads) (Cumulative)
- factor - Unknown field
- segment\_actual\_time - This is a segment time. Time taken by

the subset of the package delivery

- segment\_osrm\_time – This is the OSRM segment time. Time taken by the subset of the package delivery
- segment\_osrm\_distance – This is the OSRM distance. Distance covered by subset of the package delivery
- segment\_factor – Unknown field

## Summary

### Insights and Observations:

- A total of 14,817 distinct trips were recorded between various source and destination points during September and October of 2018.
- There were 1,504 delivery routes utilized for these trips.
- The data comprises 1,508 unique source centers and 1,481 unique destination centers.
- Among the 14,817 total distinct trips:
  - 8,908 (60%) were categorized under Carting, employing small vehicles,
  - 5,909 (40%) were categorized as FTL (Full Truck Load), facilitating quicker deliveries due to the absence of intermediate pickups or drop-offs.

### Hypothesis Test Outcomes:

- The two-sample t-test provided insights such as:
  - The average `time_taken_btwn_odstart_and_od_end` for the population equates to the average `start_scan_to_end_scan` for the population.
  - The population's average `actual_time` is lesser than the average `start_scan_to_end_scan`.

### Exploratory Data Analysis:

- The analysis reveals that cities like Mumbai (Maharashtra), Delhi, Gurgaon (Haryana), Bengaluru (Karnataka), Hyderabad (Telangana), Chennai (Tamil Nadu), Ahmedabad (Gujarat), Pune (Maharashtra), Chandigarh, and Kolkata (West Bengal) record a high number of intra-city trips.
- Regarding unequal source and destination states, the highest number of trips occur between: Delhi to Gurgaon, Gurgaon to Bengaluru, Bhiwandi/Mumbai to Pune (Maharashtra), and Sonipat to Gurgaon (Haryana).
- Numerous deliveries to airports were noted, including routes like Chennai to MAA Chennai International Airport, Pune to Pune Airport (PNQ), Kolkata to CCU West

Bengal Kolkata International Airport, and Bengaluru to BLR-Bengaluru International Airport.

- From the Bar charts and calculated tables, it's apparent that the majority of trips occur within specific cities. Moreover, regarding average distance between destinations, routes like Guwahati to Mumbai, Bengaluru to Chandigarh, Bengaluru to Delhi, and Bengaluru to Gurgaon are among the longest.

## Significant Long-Distance Routes:

- Routes with substantial distances and high trip occurrences include: Guwahati to Bhiwandi, Bengaluru to Chandigarh, Bengaluru to Delhi, Gurgaon to MAA Chennai Airport, Bhiwandi to Kolkata, Bengaluru to Kolkata, Gurgaon to Hyderabad, and Gurgaon to Kolkata.

## Multi-City Routes:

- Notable multi-city routes include: Guwahati to LakhimpurN, Jaipur to Tarnau, Guwahati to Tura, Mangalore to Udupi, Ajmer to Raipur, Mainpuri to Tilhar - each traversing through more than eight cities.

## Busiest Routes and High-Activity States:

- Delhi to Haryana emerges as the busiest route with over 400 trips. Other busy routes include Haryana to Uttar Pradesh, Chandigarh to Punjab, and Delhi to Uttar Pradesh.
- States like Maharashtra, Karnataka, Tamil Nadu, Haryana, Telangana, Gujarat, West Bengal, and Uttar Pradesh have recorded over 1,000 trips (Cells In.[173]).

## Major Traffic Warehouses:

- Warehouses experiencing significant traffic, and thus identified as busiest junctions, include: Bengaluru (Karnataka), Gurgaon (Haryana), Mumbai (Maharashtra), Hyderabad (Telangana), Delhi, Pune (Maharashtra), Chandigarh (Punjab), Chennai (Tamil Nadu), Sonipat (Haryana), Kolkata (West Bengal), Ahmedabad (Gujarat), MAA (Tamil Nadu), Jaipur (Rajasthan), Kanpur (Uttar Pradesh), Surat (Gujarat), Muzaffarpur (Bihar), FBD (Haryana), Bhopal (Madhya Pradesh), and Noida (Uttar Pradesh).

## Recommendation

- Based on the analysis, employing Carting (small vehicles) for intra-city deliveries is advisable to expedite delivery times. For long-distance or heavy load deliveries, utilizing Heavy trucks is recommended. By adhering to this strategy, delivery times



can be optimized which in turn, may enhance revenue generation according to the operational requirements.

- Enhancing connectivity in tier 2 and tier 3 cities, coupled with strategic partnerships with various e-commerce giants, can potentially boost both revenue and reputation regarding cross-border connectivity.
- Efforts can be channeled towards refining the scanning process at both the initiation and conclusion phases. By optimizing the start and end scanning times, it's plausible to align the actual delivery times closer to the OSRM estimated delivery times.

## 5. Walmart - Confidence Interval and CLT

### Project Details

 Walmart - Confidence Interval and CLT

### Learning Summary

- Analyzed customer purchase behaviors at Walmart, focusing on gender, marital status, and age during Black Friday.
- Utilized Central Limit Theorem and bootstrapping for statistical analysis.
- Found men generally spend more than women; no significant difference in spending between married and single customers.
- Identified 26-35 age group as 40% of customers, with spending reflecting overall average.
- Recommended targeted marketing strategies for various demographics to optimize revenue.

### Introduction

Walmart is an American multinational retail corporation that operates a chain of supercenters, discount departmental stores, and grocery stores from the United States. Walmart has more than 100 million customers worldwide.

#### **Business Problem**

The Management team at Walmart Inc. wants to analyze the customer purchase behavior (specifically, purchase amount) against the customer's gender and the various

other factors to help the business make better decisions. They want to understand if the spending habits differ between male and female customers: Do women spend more on Black Friday than men? (Assume 50 million customers are male and 50 million are female).

## Dataset

The company collected the transactional data of customers who purchased products from the Walmart Stores during Black Friday. The dataset has the following features:

Dataset link: [Walmart\\_data.csv](#)

The dataset have the following fields:

## Summary

- 75% male and 25% are female customers as per given sample data
- (With 95% confidence and sample size of 10000 , 500 trials.), As per confidence Interval comparison for both female purchase and male purchase data , its clear that there's no over lapping , and hence there's a good amount of difference between Male and Female Spending amounts .
- Male Customers are more likely to spend more amount than female customers .
- Average Male Spending Amount from all 100 million customers lies in Range of 9333 to 9533 as per Bootstrapping Method
- Average Female Spending Amount from all 100 million customers lies in Range of 8639 to 8826 as per Bootstrapping Method
- As per confidence Interval comparison for both Single and Married Customer's average purchase data
- There is not much difference between their average spending amounts. Married and Single Customer's spending amounts distribution are almost lies with same distribution.
- Customers from age 26-35 are 40% of all customers. and their Average Spending amount is near to overall customers average spending amount
- Age group 51-55 customers are more likely to spend more amount than all other groups
- Customers under 17 age are the least spending average amount

- As per calculations and above distribution plot, as we increase the sample size, standard error decreases, means that the average spending amount gets closer and closer to the actual mean spending amount of the all customer average spending amount.
- All city categories are having customers majorly who are living there for 1 to 2 years.
- Out of all women, 35% of the revenue coming from age group 18 to 45 and so is same for men as well.

## Recommendation

- City Category B has the highest customer base compared to C and A. Since City Category A and C customers, have the lesser spending average amount that city category B customers, more infrastructure and marketing strategies can be focused on City category A.
- There is not much significant difference between Married and Single Category Customers, no changes needs to be taken in that area.
- And there is a huge gap and difference between Male and Female spending average amounts and intervals, We can introduce special offers for particularly women like Women's day offer, or mother special or something like that.
- Age group 0-25 has the lowest spendings compared to other age groups. Since most of the 0-25 age customers would be students, more products related students / teenage / kids recommended to introduce and university/student discount can help increase the revenue from this age group.

In [ ]: