SAS Project - Bike Sharing Rider Membership Analysis

By Peng Wang
Instructor: Mr. Ar Kar Min
Metro College of Technology
2021 July 30

By analyzing the 2018 bike-sharing rides of Washington City, this project provided the business insight into evaluating the current member loyalty and market share strategy and consequently maximizing future bike-sharing sales.

"Customer satisfaction is worthless. Customer loyalty is priceless."
- Jeffrey Gitomer.

Executive Summary

The bike-sharing market was valued at USD 3 billion in 2020, and it is anticipated to reach USD 4 billion by 2026, registering a CAGR of about 6% during the forecast period (2021 – 2026).[1]

The marketing of bike-sharing is increasingly competitive. Therefore, attracting more riders to have a membership with the company is a critical factor in enhancing customer loyalty and market share.

Figure 1 clearly shows that riders with membership and without membership (casual riders) have significantly different activity patterns. Analyzing these patterns may help businesses improve their current marketing strategy and maximize future sales.

[1] https://www.mordorintelligence.com/industry-reports/bike-sharing-market

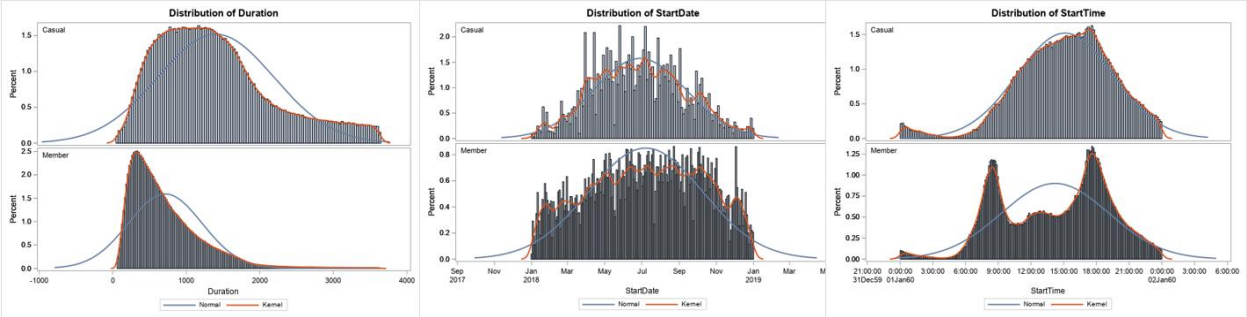


Figure 1 Different Activity Patterns between Member Types

Sixteen business questions have been reviewed and answered throughout the analysis. They are highly related to sales increasing or costs effectiveness.

Q1:How many bikes are available for sharing?	Q2: How many stations in the city?	Q3: Which are top 50 most popular start station?	Q4: What is the minimum and maximum duration? How long is the average duration?
Q5: What day is the busiest day during a week?	Q6: What time slot is the busiest during a day?	Q7: Which month is the busiest or most quiet month?	Q8: What is ratio of rider's membership?
Q9: Does Weekday has any impact on Membership?	Q10: Does Timeslot has any impact on Membership?	Q11: Does Start Station has any relationship with Membership?	Q12: Does End Station has any relationship with Membership?
Q13: Does Month has any impact on Membership?	Q14: Does duration has any relationship with Membership?	Q15: Does startDate has any relationship with Membership?	Q16: Does startTime has any relationship with Membership?

Figure 2 Questions during the Descriptive Analysis

Business Background

According to the source data, in 2018,

- there were 528 bike-sharing stations across Washington city, and
- 5387 bikes were available for the ride.
- Both member and casual riders achieved 3.5 million rides, i.e. 10K rides each day.

- The total riding duration was 790K hours, i.e. 13.5 min for each ride on average.

Member riders contributed 79 percent of the rides count but only 69 percent of the riding duration.

Member_type	Total_Duration
Casual	875343070
Member	1969958315

Member_type	Method	N	Mean
Casual		624396	1401.9
Member		2757650	714.4

Figure 3 Rides Duration, Counts, and Average Duration

Interestingly, the casual riders have an average riding duration of 1400 minutes, while member riders' average riding duration was just 714 minutes.

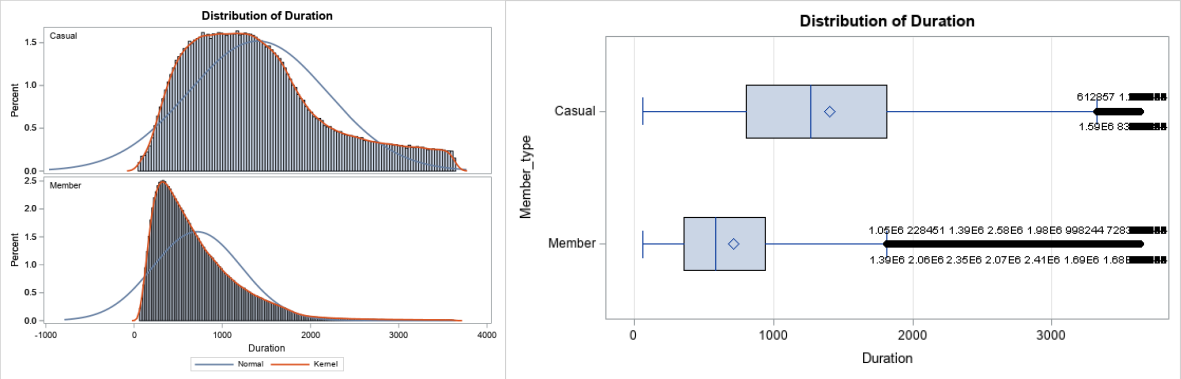


Figure 4 Histogram (left) and Boxplot (right) of Duration comparison between Member_Types

Project Objective

1. **Business Objective**: To figure out how to increase customer loyalty and improve market share and achieve more sales in the future.
2. **Technical Objective**: To predict rider's membership type by analyzing the ride-sharing records.

Methodology

1. Data Collection — Source data provided by Metro College of Technology.
2. Data Definition — 2018 Washington City bike-sharing records.
3. Data Scope — 12 .csv files, 9 variables, 3.5 million observations.

4. Software Used SAS 9
5. Statistics Methods Used
    i. Data Cleaning — Handle missing values
    ii. Feature Engineering — Handle outliers
    iii. EDA — Univariate & Bivariate Analysis, Hypothesis test
    iv. Data Visualization — Histogram, Bar Chart, Pie Chart, Boxplot
    v. Logistic Regression

Study Variables

The source data contains 12 .csv files (one for each month of 2018). All 12 files included the same variables and followed the same column structure. The total observation is 3.5 million bike-sharing rides.

There were nine attributes include the target variable "Member_Type," which has two levels: Member and Casual.

Type	Field Name	Original Field Name	Changes	Preview
#	Duration	Duration		552, 1,282, 1,265
📅	Start date	Start date		2018-01-01, 12:05:06 a.m., 2018-01-01, 12:14:30 a.m., 2018-01-01, 12:14:53 a.m.
📅	End date	End date		2018-01-01, 12:14:18 a.m., 2018-01-01, 12:35:53 a.m., 2018-01-01, 12:35:58 a.m.
#	Start station nu...	Start station number		31,104, 31,321
Abc	Start station	Start station	✎	Adams Mill & Columbia Rd NW, 15th St & Constitution Ave NW
#	End station num...	End station number		31,400, 31,321
Abc	End station	End station	✎	Georgia & New Hampshire Ave NW, 15th St & Constitution Ave NW
Abc	Bike number	Bike number		W00886, W01435, W21242
Abc	Member type	Member type		Member, Casual

Figure 5 A Quickview of Original Variables

Challenges:

Start dates: they are in mixing date&time format. For analysis, I need to convert (separate) it into different columns: Month, StartDate, StartTime, Weekday, StartTimeSlot (e.g. Rush Hour, Early Morning, Afternoon).

End dates: Dataset has the Duration and Start date; therefore, it can calculate the End Date. Moreover, the critical attribute for my analysis is Start Date, as it represents the "moment of the business need." Therefore, it is unnecessary to keep the End date, so I dropped it.

Start station / End station: they are text values and accompanied with Start station number / End station number. For my project scope, I only need Start/End station numbers. Therefore I dropped them.

Figure 6 shows the final dataset ready for SAS analysis:

Month	StartDate	Weekday	StartTimeSlot	StartTime	Duration	Start station number	End station number	Bike number	Member type
1	2018-01-17	Wednesday	Rush Hour	08:14:58	307	31606	31606	W00076	Member
1	2018-01-17	Wednesday	Rush Hour	08:15:04	657	31248	31290	W01224	Member
1	2018-01-17	Wednesday	Rush Hour	08:15:28	501	31611	31623	W00167	Member
2	2018-02-26	Monday	Afternoon	16:40:48	542	31209	31610	W01327	Member
2	2018-02-26	Monday	Afternoon	16:40:50	1,456	31216	31628	W22439	Member
2	2018-02-26	Monday	Afternoon	16:40:50	543	31209	31610	W20752	Member
2	2018-02-26	Monday	Afternoon	16:40:55	195	31048	31045	W22595	Member
2	2018-02-26	Monday	Afternoon	16:41:01	353	31288	31248	W20370	Member
2	2018-02-26	Monday	Afternoon	16:41:10	341	31231	31232	W23117	Member
2	2018-02-26	Monday	Afternoon	16:41:20	851	31625	31614	W20129	Member
2	2018-02-26	Monday	Afternoon	16:41:25	1,119	31206	31623	W22890	Member

Figure 6 Example of Finalized Dataset (Tableau)

Feature Engineering:

I finally selected the below list of Predictor Variables to model the Target Variable.

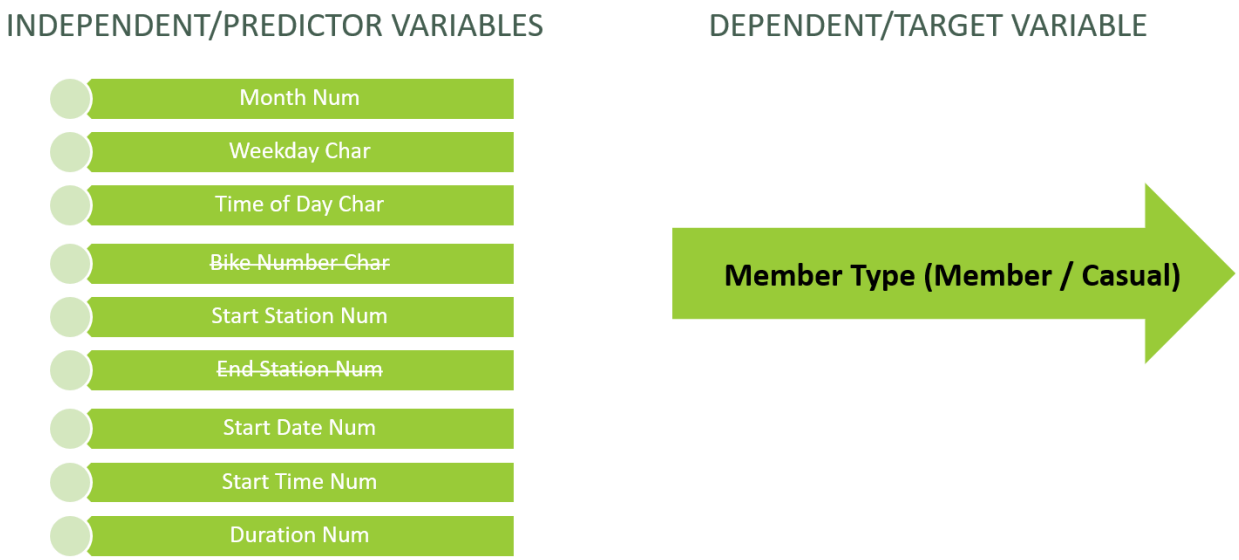


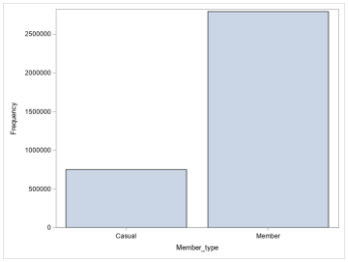
Figure 7 Conceptual Framework

I have further removed the Bike Number and End Station Number variables, as Bike Number is just an ID that has no relationship with the membership types, and End Station Number is logically duplicated with the Start Station Number – for the same station, it might be the start station in a record but the end station in another. Moreover, same as Start Date or Start Time, I prefer to keep the Start Station in the list as it represents the moment of business needs.

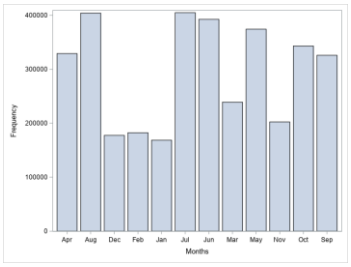
Univariate Analysis

A few interesting findings:

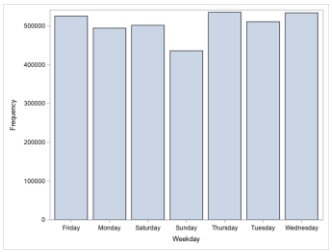
1. **20/80 Rule applied** here, too, in terms of the business activities between Members and Casual Riders.

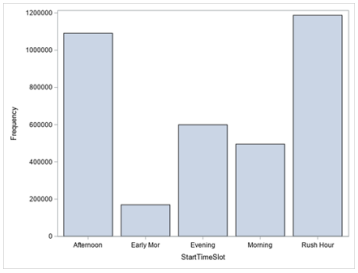


2. A typical outdoor business scenario in the Northern Hemisphere cities: business booms in Spring, Summer is the peak season, starts falling in Fall, then finally becomes quiet during the Winter.



3. **Wednesday, Thursday and Friday** took the top 3 days of rides count. Surprisingly it's not Saturday and Sunday (which is the pattern of using bike-sharing in my family).





4. The use of bike-sharing has similar counts between Afternoon and Rush Hour.

5. **The usage gap** between the top 50 popular stations and the top 50 quiet stations is HUGE. For example, the top 1 popular station got 63493 visits to compare 12 trips for the most modest station.

The top 50 quietest stations - Do we really need them?

Start_station	COUNT	PERCENT
32228	12	.000338726
31819	15	.000423408
32030	16	.000451635
32042	22	.000620998
32407	24	.000677452
32405	40	.001129087
31815	42	.001185542
31822	42	.001185542

6. **90** stations have less than **ONE visit** per day. So does a business need to spend costs to maintain them?

7. The outlier involved in Duration has a small portion but impacts a lot.

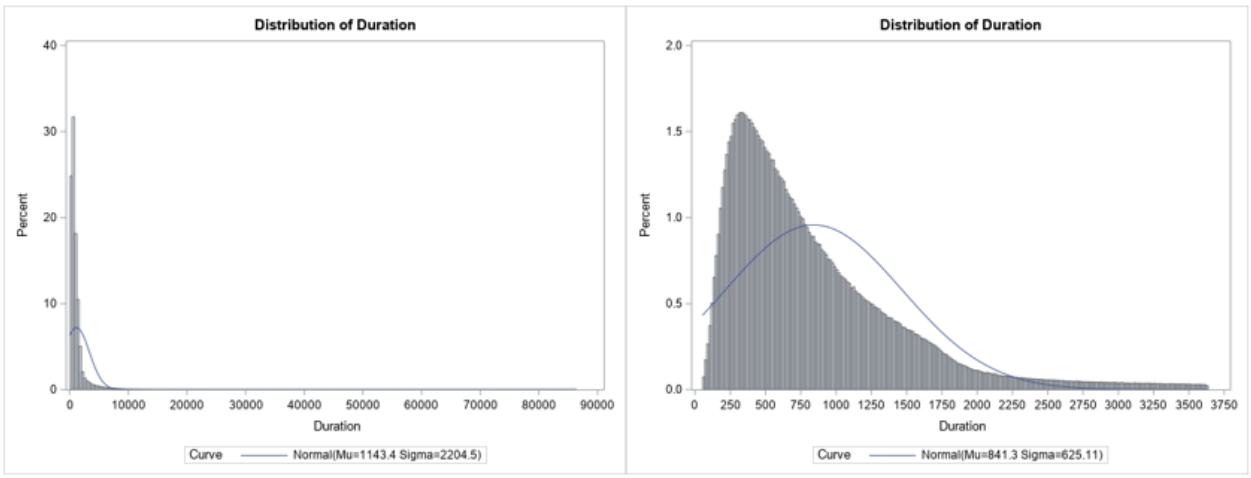


Figure 8 Duration Outlier Handling Before (left) vs. After (right)

The logical principle followed in handling outliers[2]:

- ✓ Anything < Q1 – 3*IQR and
- ✓ Anything > Q3 + 3*IQR will be removed, provided
- ✓ They are less than 10% of the total records.

Total 160,638 observations (4.5%) had been removed.

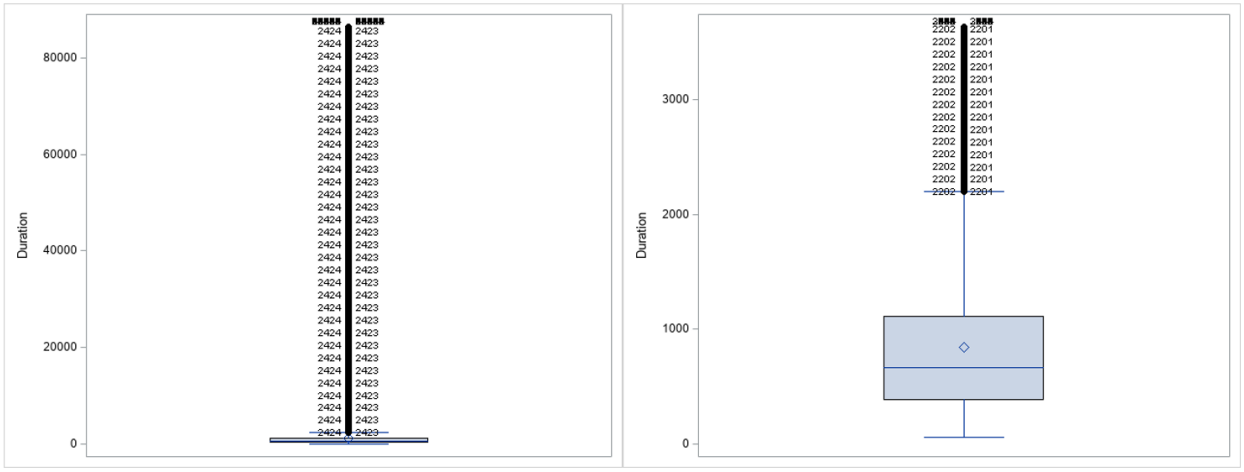


Figure 9 Outlier Removed Before (left) vs. After (right)

Bivariate Analysis

- Chi-square Tests
  - ✓ %chi_sq(weekday,member_type);

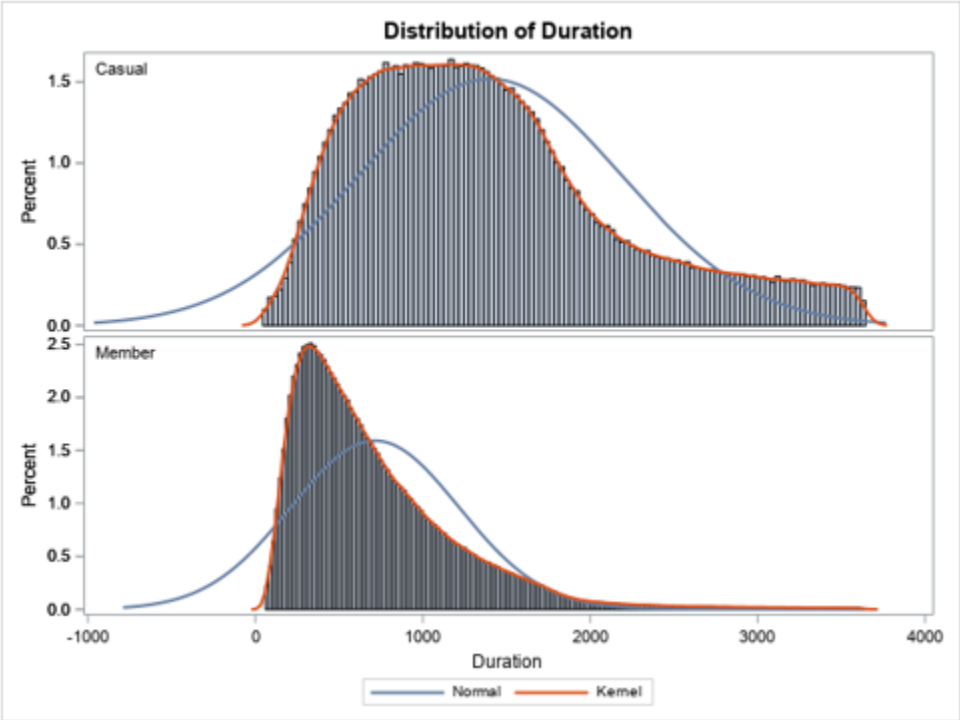
[2] https://www.listendata.com/2014/10/identify-and-remove-outliers-with-sas.html

- ✓ %chi_sq(StartTimeSlot,member_type);
- ✓ %chi_sq(start_station,member_type);
- ✓ %chi_sq(end_station,member_type);
- ✓ %chi_sq(months,member_type);

- T-Tests
  - ✓ %ttest(duration,member_type);
  - ✓ %ttest(startDate,member_type);
  - ✓ %ttest(startTime,member_type);
- Anova Tests
  - ✓ %anova(weekday);
  - ✓ %anova(StartTimeSlot);
- Correlation Tests
  - The dataset has just one continuous variable - Duration. The rest two numerical variables are date and time across the whole year, which is not related to my analysis scope.

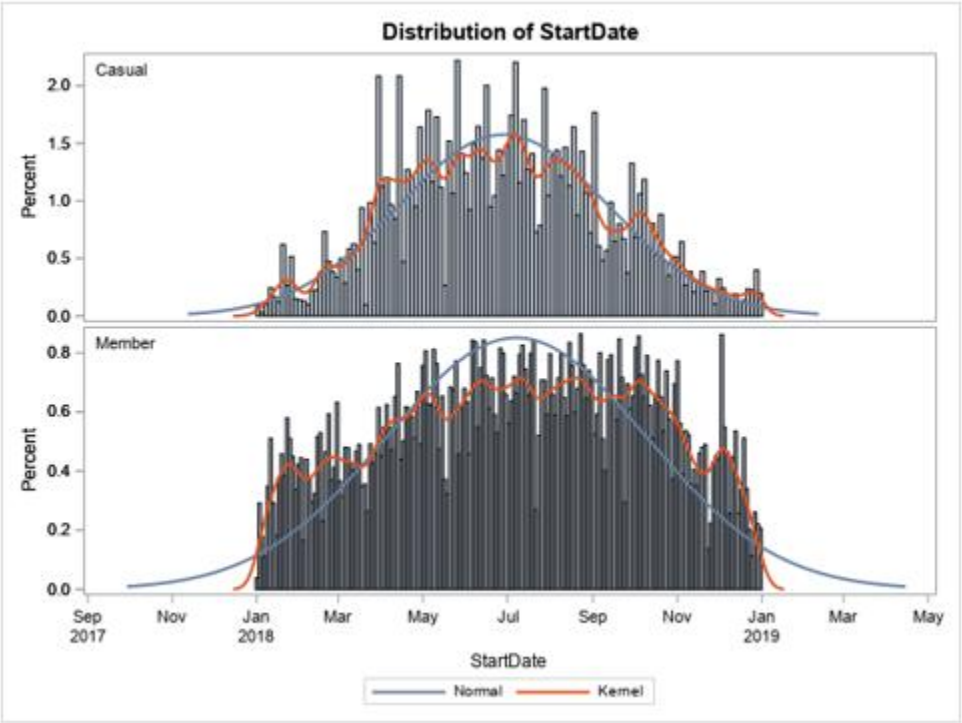
Interesting Findings:

1. Member Riders ride short Duration; Casual Riders ride much more prolonged than Members. Really? Members pay annual fees for a short duration, but casual riders pay expensive rates and keep the bikes longer? Why?



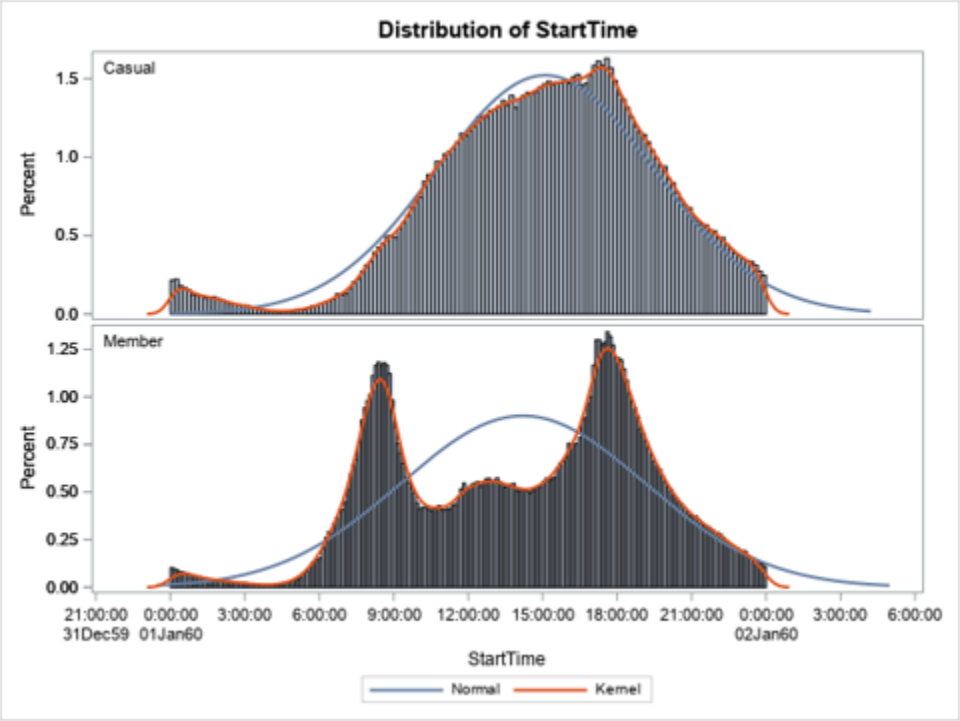
2. Members contributed a steady business during the year; however, casual riders mainly had more activities during the Summers. Therefore, it looks like members riders have consistent demand or reasons for using the bike-sharing services, while casual users ride

the bikes for seasonal-oriented activities such as sightseeing.



3. The peak time for members is Rush Hours – 7-9 am and 5-7 pm. However, they were reticent between 9 am – 5 pm. Why? I guess (obviously) they are working class. In contrast, casual riders were quiet during the rush hours, and start booming sharply around 9 am until reach the peak time of 8 pm, even having more activities during the midnight and earlier morning, compare with members. Why? I guess they are visitors,

retired people, or young students.



Hypothesis Conclusions:

The P-values shown in the previous Bivariate Analysis proved that the below variables have a P-value of less than 5% significant level. Therefore we reject the Null Hypothesis.

In other words, there is a significant statistic relationship between the below variables and Membership Type:

- Weekday
- Time of the Day
- Months
- Duration
- Station

Conclusions & Recommendations

Conclusions	Recommendations
1. 90 stations have less than one visit per day. 2. 79% of rides are from members, while 21% of rides are from casual users.	1. Cancel or relocate those 90 stations that have significantly fewer visits to save the operations costs.

3. However, casual riders have almost twice the average riding time than members. 4. Members contributed a steady business across the year, while casual riders brought more business during the summer. 5. Members have two peak times during a day matching the regular city traffic rush hours, while casual users increased throughout the day, especially from noon to evening.	2. For 21% rides from casual users, promoting competitive pricing and membership benefits to attract these riders convert to membership, therefore enhancing the marketing share. 3. Members seem to use bikes for work or shopping too much instead of enjoying the "RIDING." Suggest to promote flexible riding time or free-ride day to members, encourage them to ride more at a different time or during another day.

Appendix

```
2.3 Handling Missing Data
-------------------------
* Make a copy;
data pw.bs(label='Copy of Original');
      set pw.bike_sharing;
run;
* Count Missing Value for all columns
https://blogs.sas.com/content/iml/2011/09/19/count-the-number-of-missing-values-for-
each-variable.html
Create a format to group missing and nonmissing
https://www.analyticsvidhya.com/blog/2014/11/sas-proc-format-guide/;
proc format library=pw.miss_fmt;* Save UDF to the same library for future use;
      value $missfmt ' '='Missing' other='Not Missing';
      value  missfmt  . ='Missing' other='Not Missing';
run;
options fmtsearch=(pw.miss_fmt);
* List of missing/non-missing, by variable;
title "List of Missing/Non-Missing by Variable";
proc freq data=pw.bs;
      format _CHAR_ $missfmt.;
      tables _CHAR_ / missing missprint nocum /*nopercent*/;
      format _NUMERIC_ missfmt.;
      tables _NUMERIC_ / missing missprint nocum /*nopercent*/;
run;title;
* Result: NO MISSING VALUES.
2.4 Feature Engineering
-----------------------
;
data pw.bs1;
 set pw.bs;
  * Convert Start/End Station Number from numerical to categorical: PUT()
```

```
https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/lepg/n04koei84kuaodn1g21eyx4btome.htm
;
  Start_station = put(Start_station_number,best5.);
  End_station = put(End_station_number,best5.);
  drop Start_station_number End_station_number;
  * Convert Month from Numerical to categorical: IF...ELSE IF;
  if Month=1 then Months="Jan";
   else if Month=2 then Months="Feb";
   else if Month=3 then Months="Mar";
   else if Month=4 then Months="Apr";
   else if Month=5 then Months="May";
   else if Month=6 then Months="Jun";
   else if Month=7 then Months="Jul";
   else if Month=8 then Months="Aug";
   else if Month=9 then Months="Sep";
   else if Month=10 then Months="Oct";
   else if Month=11 then Months="Nov";
   else Months="Dec";
  drop Month;

Run;


* https://support.sas.com/kb/36/898.html
Create a Macro procedure receives variable as parameter and return the unique values (levels)
of that variable;
%macro levels (var);
       ods select nlevels; *Use the NLEVELS option with the ODS SELECT statement to capture the
                                          number of levels for a variable;
       proc freq data=pw.bs1 nlevels;
       tables &var;
       title "Number of unique level for &var";
       run; title;
%mend levels;
*Q1:How many bikes are available for sharing?;
%levels(Bike_number);
* ANSWER: 5387 BIKES ARE AVAILABLE FOR RIDE SHARING.
Q2: How nmay stations in the city?;
%levels(Start_station);
%levels(End_station);

* ANSWER: THERE ARE 528 BIKE SHARING STATIONS IN THE CITY.
Q3: Which are top 50 most popular start station?
https://communities.sas.com/t5/Statistical-Procedures/How-to-find-3-most-frequently-occurring-values/td-p/204661
;
proc freq data=pw.bs1 order=freq;
tables Start_station / noprint out=station;
```

```
run;
proc print data=station (obs=50) noobs;
title "The top 50 popular stations";
run; title;
* ANSWER: Top 50 popular start stations are found. Interestingly, if look at their percentages,
these stations
are not significantly above others. For example, the most popular station only has 1.79%.
In another word, there's no a single station is significantly busy or popular than others.
Moreover, if look at top 50 quietest stations:
;
proc freq data=pw.bs1;
tables Start_station / noprint out=station;
run;
proc sort data=station;
by count;
run;
proc print data=station (obs=50) noobs;
title "The top 50 quietest stations - Do we really need them?";
run; title;
* we found out that for the whole year of 2018, these stations had been visited from
the minimum 3 times per 1 million visits to
the maximum 40 times per 1 million visits.
For such less usage, the business question might be: DO WE REALLY NEED TO KEEP MAINTAINING THESE
STATIONS?
Further more, I listed start stations with count < 365, i.e., all stations had less than 1 visit
per day during 2018;
proc print data=station;
       where count < 365;
       title "Stations less than 1 visit per day - Do we really need them?";

run; title;
5.     Machine Learning (Modeling);
proc logistic data = pw.bs3 desc;
class months weekday starttimeslot start_station;
model member_type = duration months weekday starttimeslot start_station;
output out = model1 p = pred_prob lower = low upper = upp;
run;
quit;
proc logistic data = pw.bs3 desc;
class weekday starttimeslot ;
model member_type = duration weekday starttimeslot;
output out = model2 p = pred_prob;
run;
quit;
PROC LOGISTIC DATA = pw.bs3;
 CLASS starttimeslot;
 MODEL member_type (EVENT="Member") = starttimeslot /CLODDS =PL;
RUN;
PROC LOGISTIC DATA = pw.bs3;
 CLASS starttimeslot weekday;
 MODEL member_type (EVENT="Member") = duration weekday starttimeslot /CLODDS =PL;
RUN;

(The End)
```