SAS Project_
Bike Sharing Rider Membership Predict

BY PENG WANG

INSTRUCTOR MR. AR KAR MIN

METRO COLLEGE OF TECHNOLOGY

2021 JULY 30

Presentation Outlines

Executive Summary

1. This project analyzed the dataset that contains 2018 bike sharing details for the Washington city.

2. Throughout the data analysis, I answered 16 business questions closely related with the factors that impact on increasing marketing share and decreasing operation costs of the bike sharing business.

3. Based on the rider's activities patterns to predict rider's membership type, to help business make the right decision on the marketing plan and pricing strategy.

Business Background

1. Bike Sharing marketing trend

2. Bike Sharing market (2018) in the Washington city
   - 3.5 million rides each year; 10K+ each day!
   - 5387 bikes to share;
   - 528 bike sharing stations;

3. Business challenges
   - 90 stations have been visited less than 1 time per day;
   - Casual users have longer average riding duration than members.

Project Objective

Technical Objective

To predict rider's membership type by analyzing ride sharing records.

Business Objective

To increase customer loyalty and market share therefore achieve more sales.

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

INDEPENDENT/PREDICTOR VARIABLES

- Month Num
- Weekday Char
- Time of Day Char
- ~~Bike Number Char~~
- Start Station Num
- ~~End Station Num~~
- Start Date Num
- Start Time Num
- Duration Num

DEPENDENT/TARGET VARIABLE

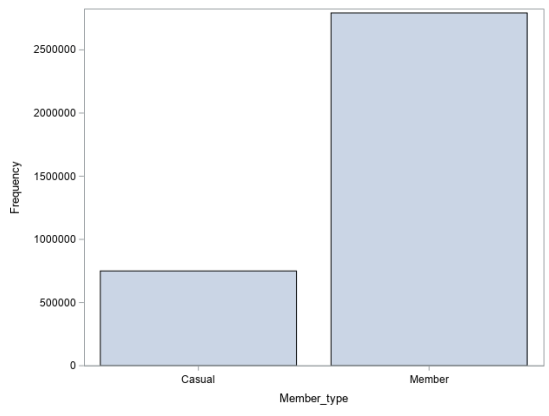
Member Type (Member / Casual)

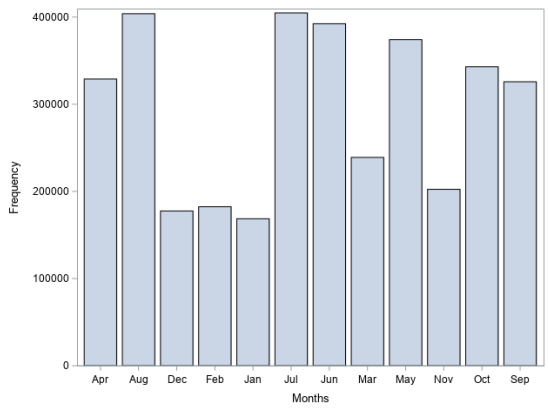
Descriptive Analysis

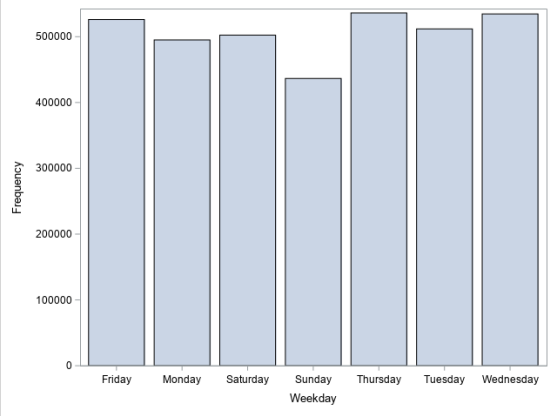
1. Univariate
2. Bivariate

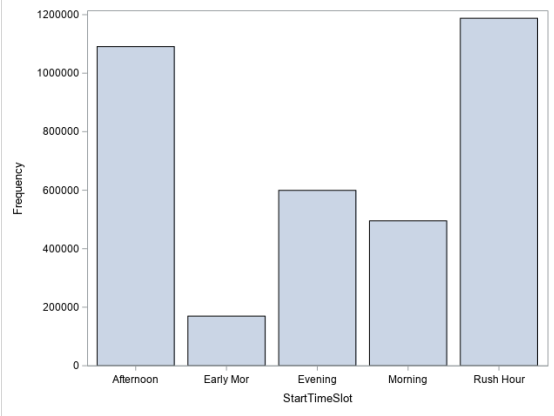
Q1:How many bikes are available for sharing?	Q2: How many stations in the city?	Q3: Which are top 50 most popular start station?	Q4: What is the minimum and maximum duration? How long is the average duration?
Q5: What day is the busiest day during a week?	Q6: What time slot is the busiest during a day?	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?

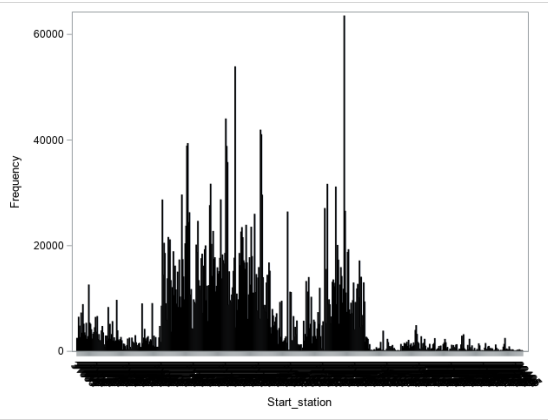
Univariate Analysis

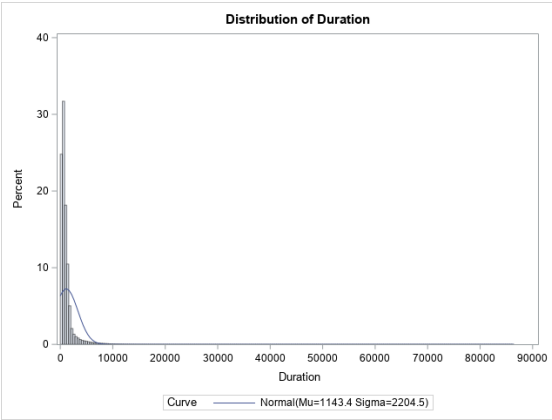


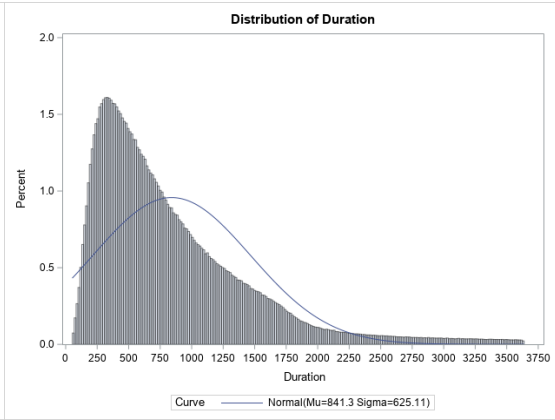












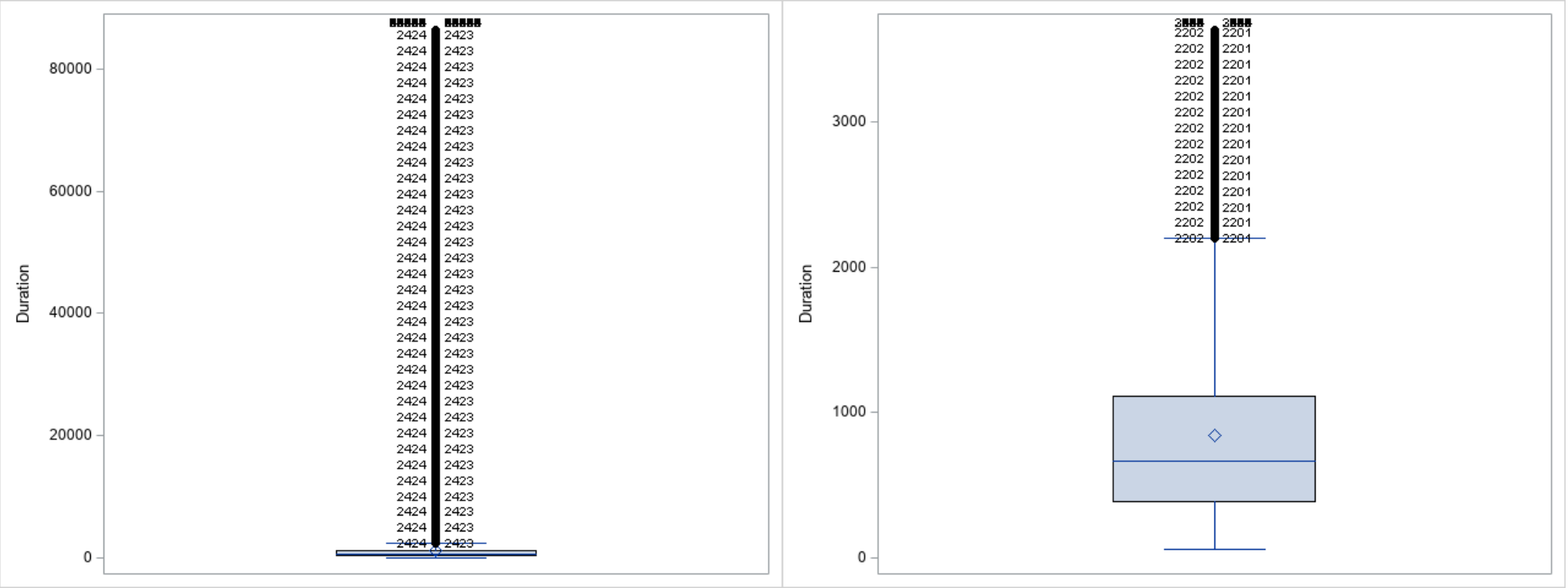
Univariate Analysis – Outlier Handling

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

Logical Principle followed in handling outliers:

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

%***outliers***(input=pw.bs2, var=Duration, output=pw.bs3);
Total 160,638 observations (4.5%) had been removed.;



Bivariate Analysis

Chi-square Tests:

%***chi_sq***(weekday,member_type);
%***chi_sq***(StartTimeSlot,member_type);
%***chi_sq***(start_station,member_type);
%***chi_sq***(end_station,member_type);
%***chi_sq***(months,member_type);

Anova Tests:

%***anova***(weekday);
%***anova***(StartTimeSlot);

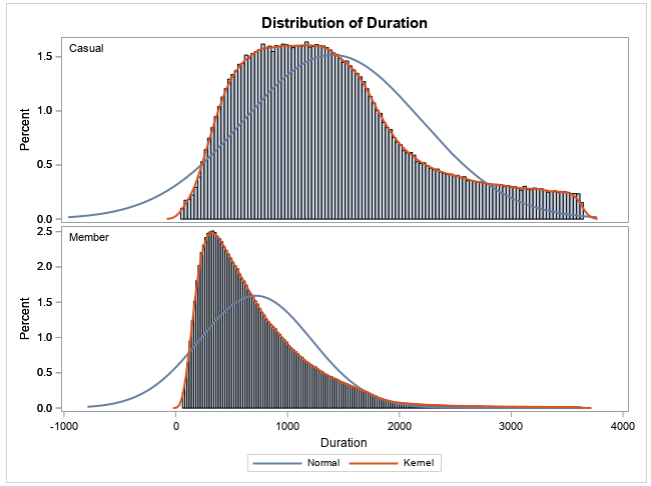
T-Tests:

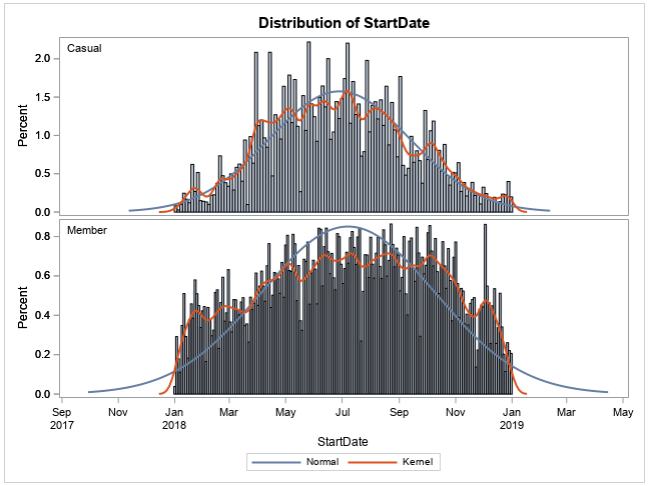
%***ttest***(duration,member_type);
%***ttest***(startDate,member_type);
%***ttest***(startTime,member_type);

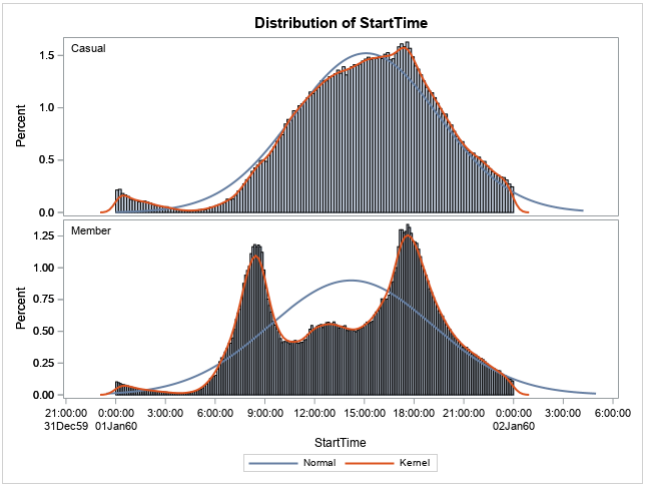
Correlation Tests:

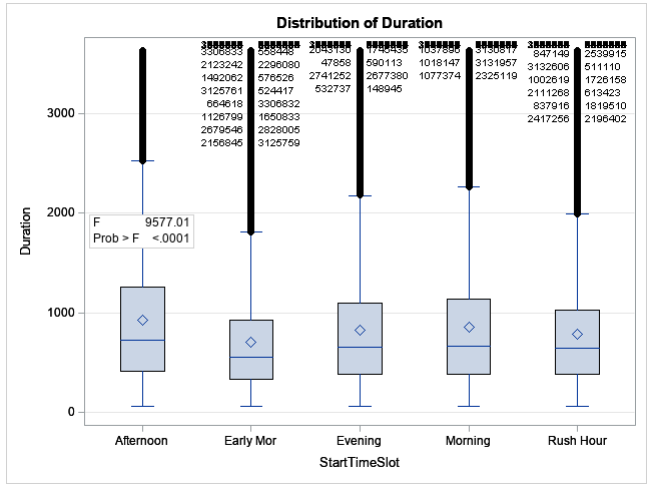
The dataset has just one continuous variable - Duration. The rest two numerical variables are date and time across the whole year, which is not related with my analysis scope.

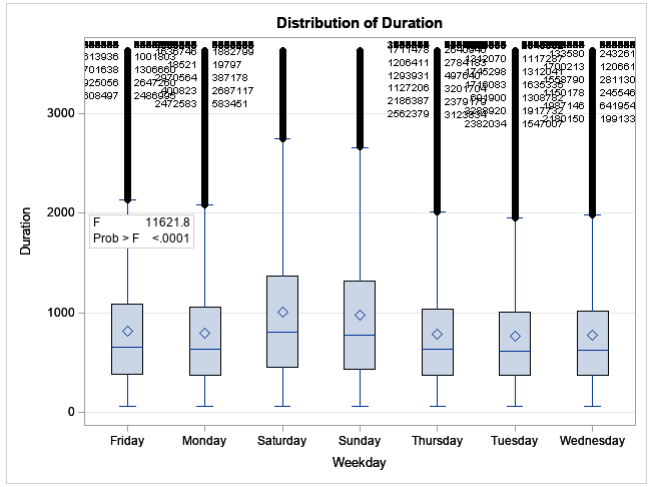
Bivariate Analysis – cont.











Inferential Statistics

1. Hypothesis Testing
2. Inference

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

In another word, there is significant statistic relationship between below variables and Membership Type:

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

Predictive Analysis

Tuning a model for 3.5m observations is challenging. It takes about 3 hours for each run, which is not something I was aware. Therefore, it leaves me limited capabilities to tune model with different parameters.

```
331  PROC LOGISTIC DATA = pw.bs3;
332   CLASS starttimeslot weekday;
333   MODEL member_type (EVENT="Member") = duration weekday starttimeslot /CLODDS =PL;
334  RUN;

NOTE: PROC LOGISTIC is modeling the probability that Member_type='Member'.
NOTE: Convergence criterion (GCONV=1E-8) satisfied.
NOTE: There were 3382046 observations read from the data set PW.BS3.
NOTE: PROCEDURE LOGISTIC used (Total process time):
      real time           3:34.40
      cpu time            3:11.84
```

However if looking at the Percent Concordant rate and the Area Under Curve (C Statistics), the model shows the Fair result.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	80.4	Somers' D	0.608
Percent Discordant	19.6	Gamma	0.608
Percent Tied	0.0	Tau-a	0.183
Pairs	1.7218656E12	c	0.804

Logistic Models

MODEL WITH ALL KEY VARIABLES

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	84.1	Somers' D	0.683
Percent Discordant	15.9	Gamma	0.683
Percent Tied	0.0	Tau-a	0.206
Pairs	1.7218656E12	c	0.841

MODEL WITH SELECTED VARIABLES

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	48.5	Somers' D	0.216
Percent Discordant	26.9	Gamma	0.287
Percent Tied	24.6	Tau-a	0.065
Pairs	1.7218656E12	c	0.608

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	80.4	Somers' D	0.608
Percent Discordant	19.6	Gamma	0.608
Percent Tied	0.0	Tau-a	0.183
Pairs	1.7218656E12	c	0.804

Conclusions & Recommendations

Conclusions:

1. 90 stations have less than 1 visit per day.

2. 79% rides are from members while 21% rides are from casual users.

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 2 peak times during a day which are matching the regular city traffic rush hours; while casual users increased throughout the whole day, especially from noon to evening.

Recommendations:

1. Cancel or relocate those 90 stations who have extreme less visits to save the operations costs.

2. For 21% rides from casual users, suggest to promote competitive pricing and membership benefits to attract these riders convert to membership therefore enhance the marketing share.

3. Members seems use bikes for work or shopping too much instead of really enjoy the “RIDING”. Suggest to promote flexible riding time or free-ride day to members, encourage them to ride more on different time or during the different day.

Peng Wang

Data Analyst | SQL | Python | R | SAS | Tableau | PowerBI | Machine Learning

Talks about #datascience, #dataanalytics, #knowledgeshare, #criticalthinking, and #inspirationalleader

Thank You and Keep in Connect.

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.
```

Appendix

```
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;
```

Appendix

```
* 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.
```

Appendix

```
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;
```

Appendix

```
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;
```