Business Process Analysis

For

Bike Share Toronto Ridership

Submitted to:

Dr. Valarie Bell

Department of ITDS

University of North Texas - UNT

Prepared by:

Samuel Castilla

Satvika Marrapu

Suhail Bari

April, 2020.

Table of Contents

[Executive Summary 4](#_Toc36560068)

[1.0 Improvement Opportunity: Define Phase 4](#_Toc36560069)

[1.1 Problem Statement/Discussion of the process being examined 5](#_Toc36560070)

[1.2 Identification of key measures used to evaluate the success of your project 5](#_Toc36560071)

[1.3 Discussion of project scope 5](#_Toc36560072)

[2.0 Current State of the Process: Measure Phase 5](#_Toc36560073)

[2.1 Current Performance Level 5](#_Toc36560074)

[3.0 Analysis and Findings: The Analyze Phase 15](#_Toc36560075)

[4.0        Recommendations: The Improve Phase: 17](#_Toc36560076)

[5.0        Monitoring and Controlling 18](#_Toc36560077)

[6.0 Conclusion 18](#_Toc36560078)

[Work Cited 19](#_Toc36560079)

[Appendix A. 20](#_Toc36560080)

[Statistical Analysis Plan. 20](#_Toc36560081)

**Table of figures**

[Figure 1. Stations with no ID Information. 6](#_Toc36563631)

[Figure 2. Assigned ID’s to Stations. 7](#_Toc36563632)

[Figure 3. Null Value Excluded. 7](#_Toc36563633)

[Figure 4. Data Integration with CDM. 8](#_Toc36563634)

[Figure 5. Trips per Year. 9](#_Toc36563635)

[Figure 6. Trips per Quarter / Year. 10](#_Toc36563636)

[Figure 7. Time Series Plot. 10](#_Toc36563637)

[Figure 8. Time Series Plot Not Outliers 11](#_Toc36563638)

[Figure 9. Time Series Plot Linear Regression Tendency. 11](#_Toc36563639)

[Figure 10. Time Series Plot Polynomial Regression Tendency. 12](#_Toc36563640)

[Figure 11. Time Series Plot from SPSS. 12](#_Toc36563641)

[Figure 12. Time Series Decomposition. 13](#_Toc36563642)

[Figure 13. Time Series Decomposition Model and Factors. 13](#_Toc36563643)

[Figure 14. Final ARIMA Model. 14](#_Toc36563644)

[Figure 15. ARIMA Model Statistics. 14](#_Toc36563645)

[Figure 16. Residual ACF and PACF Plots 14](#_Toc36563646)

[Figure 17. Most Used Routs. 15](#_Toc36563647)

[Figure 18. Least Used Routs. 16](#_Toc36563648)

[Figure 19. Maximum Commute To/From Station Wise. 16](#_Toc36563649)

[Figure 20. Time Based Casual Member Usage. 17](#_Toc36563650)

**Contents**

# Executive Summary

This project is about the bike share service in Toronto, Canada. Our analysis is based on the second-hand data extracted from Kaggle.com, and it includes the years 2017 and 2018.

The project follows the DMAIC (Define, Measure, Analyze, Improve and Control) cycle, that is the core tool for six sigma.

The Define phase establishes that the project purpose is to increase business revenue alongside bringing in non-users and converting casual members to annual members. Also, it defines Toronto city as the scope of the project.

The current state of the process, or Measure phase, describes the dataset and the preprocessing tasks. The accuracy, completeness, timeliness, believability, and interpretability were demonstrated. In the data cleaning we used MS Excel to handle missing values and misspelling station names. For 2017\_Q4 (quarter 4 for 2017) was found 1 null value out of 363,405 records that was excluded from the dataset. About data integration, all the eight data files where integrated in only one large CSV file using the CMD tool. The phase of data processing was the longest task in this measure phase.

To make the required measurements, we used Tableau and SPSS as data analysis tools. The linear regression and polynomial regression models were not statistically significant to explain the data, that shows a stationary behavior. Therefore, we built an ARIMA model that explains the data in 86.9%, which is statistically significant.

The Analyze phase shows the usage of the routes determining the most and least used routes. Also, it established the casual and member user behavior with respect to time-based usage.

The Improve phase determines four recommendations related with Reviews of the Customers, Marketing and Information, Identifying Service Gaps, and Membership for the customers.

Finally, the Control phase establishes the best way to ensure the success in the implementation of the recommendations.

# 1.0 Improvement Opportunity: Define Phase

To increase ridership by at least 10% by 2019. The proposed deliverable will be an increase in non-member ridership by analyzing data against routes, type of users and usage (number of seconds).

The purpose of this project is to increase business revenue alongside bringing in non-users and converting casual members to annual members.

We will examine the data and find out the most travelled routes, the least travelled routes, average, to find the important shortcomings of the business and improve customer satisfaction.

## 1.1 Problem Statement/Discussion of the process being examined

In this project, we have data for the bike share program as per the data dictionary. In order to grow the business, we need to analyze the data to find the suitable way to increase customer base.

## 1.2 Identification of key measures used to evaluate the success of your project

* Analyzing the most used routes and the least used routes and its seasonal variations.
* Making the most used routes most accessible.
* Distinguish between casual users and annual members.

## 1.3 Discussion of project scope

The scope of the project is currently limited to Toronto city only.

# 2.0 Current State of the Process: Measure Phase

## 2.1 Current Performance Level

The data collected in this project comes from the online source Kaggle. This second-hand dataset provides anonymous information about the bike share ridership. It includes information from 2017 and 2018 and is divided in quartiles for each year. It has 3,415,324 records in total. The dimensions of the Bike Share dataset are Trip start date and time, Trip end date and time, Trip duration, Trip start station, Trip end station, and User type.

This dataset is from Toronto Parking Authority, published on https://open.toronto.ca/dataset/bike-share-toronto-ridership-data/. The data is licensed under: Open Government License – Toronto”. However, the source of this secondary dataset is <https://www.kaggle.com/jackywang529/toronto-bikeshare-data/data>.

The following is the Data dictionary:

* trip\_id -- A unique ID created for each trip
* trip\_start \_time -- The start date and time of the trip
* trip\_stop\_time -- The end date and time of the trip
* trip\_duration\_seconds -- The time duration of the trip in seconds
* from\_station\_id -- A unique id for each station
* to\_station\_id -- A unique id for each station
* from\_station\_name -- Name of the start station
* to\_station\_name -- Name of the end station
* user\_type -- Types of users

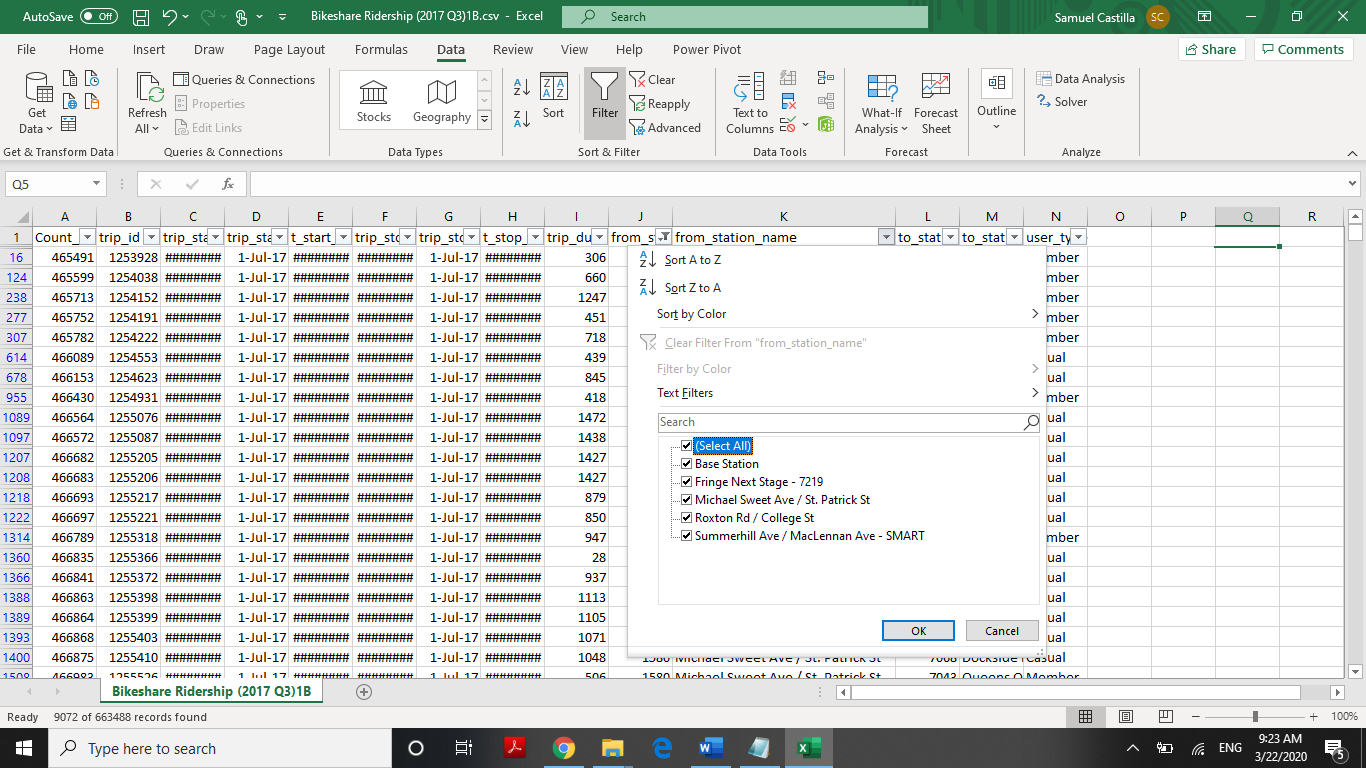
Data Preprocessing is presented as follows:

* Data quality issues:
  + Accuracy: the dataset accurately represents the behavior of bike share ridership in Toronto.
  + Completeness: the files include all data about the ridership for 2017 and 2018. Also, the data has values for all fields in it.
  + Timeliness: the data was collected in the correspondent periods, so it is plenty useful.
  + Believability: the data can be trusted, due to it comes from Toronto Parking Authority that is an official source. Also, all values make sense.
  + Interpretability: the data shows the information described in the variable head. Likewise, the values are accurate.
* Data cleaning:

The dataset for 2017 in the quarter 3 and 4, included missing values for the dimensions “from\_station\_id” and “to\_station\_id”. To handle these missing values, we used the excel function VLOOKUP to find the ID Station information in the rest of the dataset either 2017 or 2018.

However, we found another issue, there were so many misspelling station names (i.e. extra periods, one more letter, etc) for the same station. For this reason, we built a file called Station ID\_1B.cvs, that includes all the station names and ID’s from the dataset we have.

After the previous analysis, we did not find any information about five (5) station ID’s in the StationID\_1B.cvs file we built:



### Figure 1. Stations with no ID Information.

To solve this situation, we created unique ID’s for these stations. As a result, we filled out all missing values for the station ID fields in the dataset:

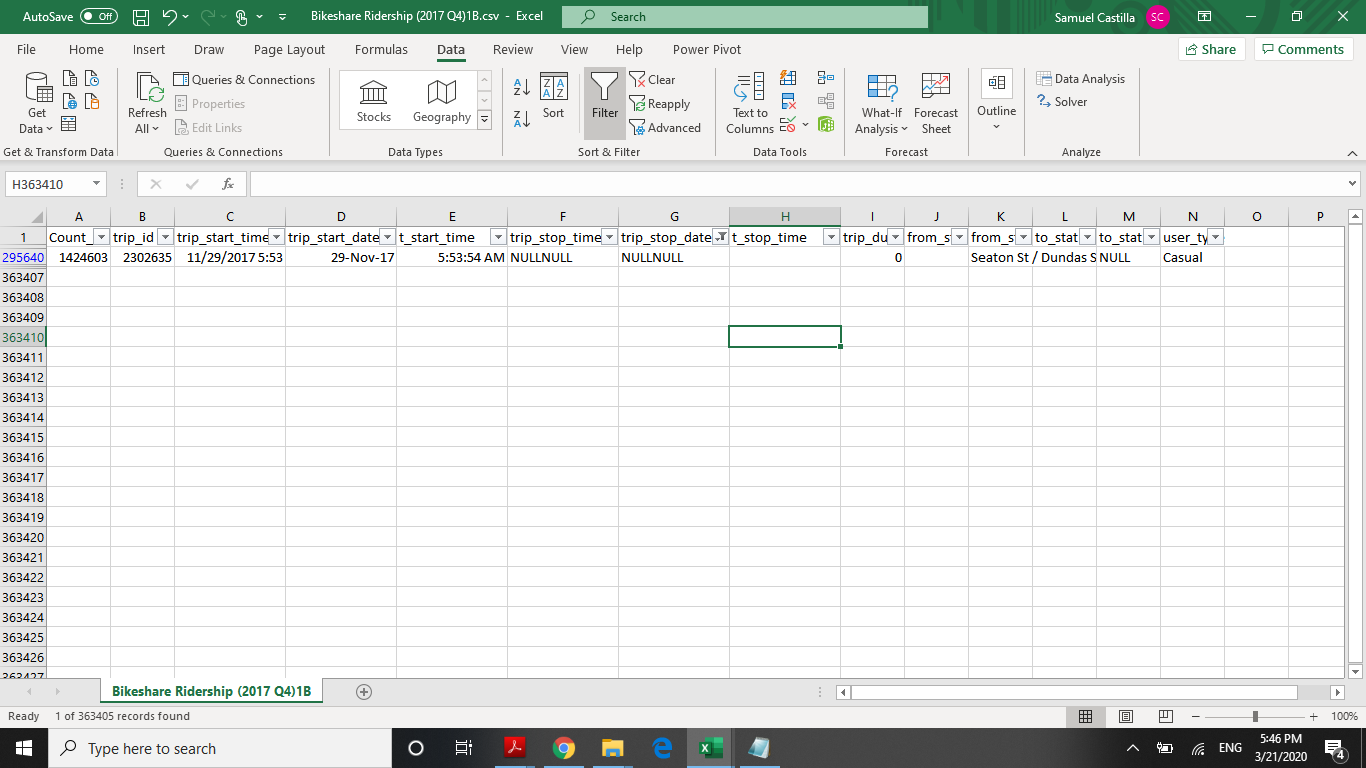


### Figure 2. Assigned ID’s to Stations.

Data for 2018 does not include any missing values, and at this time we have not identified any noisy data.

Null Values:

About null values, we found only one value for 2017\_Q4 (quarter 4 for 2017). This is 1 out 1 out of 363405 records in this period of time, for this reason the team decided to exclude it from the dataset.



### Figure 3. Null Value Excluded.

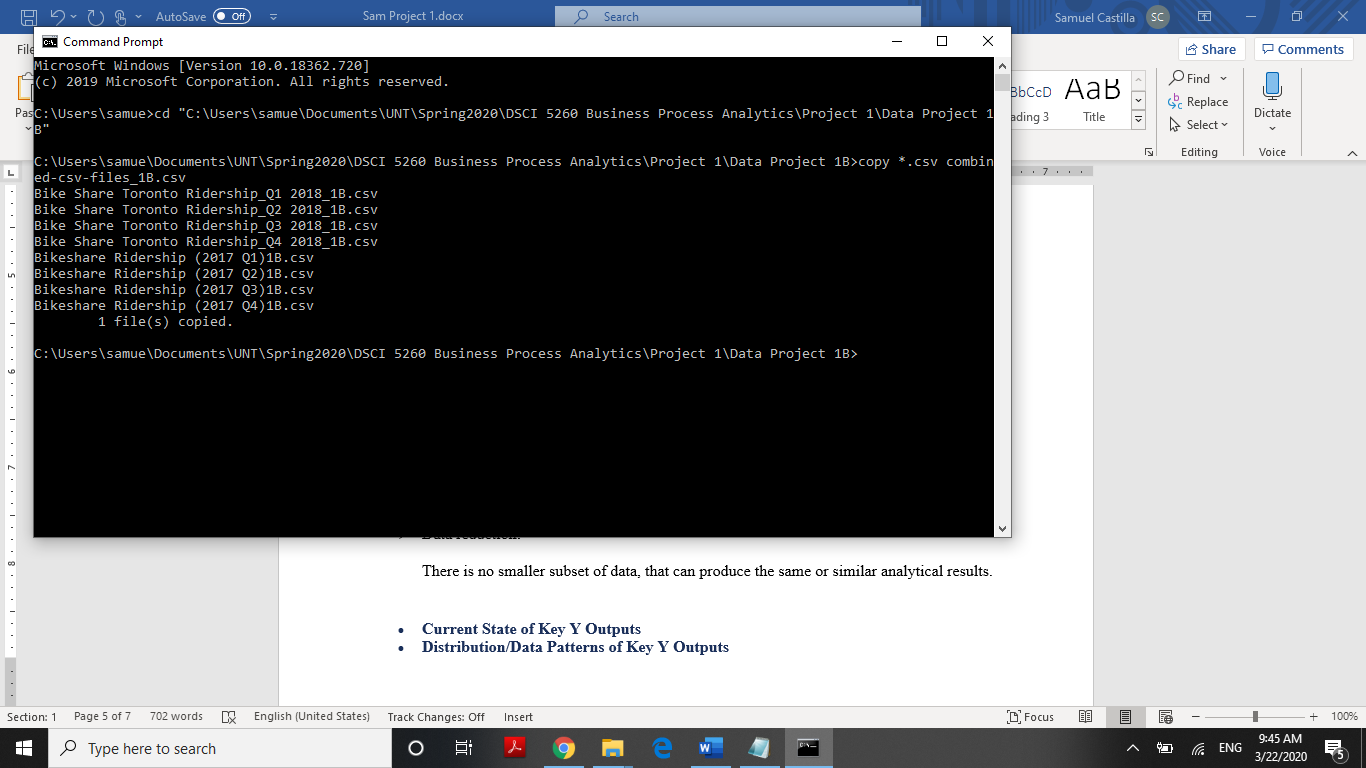
* Data Integration:

As we mentioned above, the dataset is originally divided in 8 files according with each quarter for each year 2017 and 2018. To integrate this data, we did the following tasks:

* Split the variable “trip\_start\_time” in two columns to have the date and time separately. The resultant variables are “trip\_start\_date” and “t\_start\_time”.
* Split the variable “trip\_stop\_time” in two columns to have the date and time separately. The resultant variables are “trip\_stop\_date” and “t\_stop\_time”.
* Include an index variable called “Count\_Index”. This dimension is for analysis purposes.
* Columns in the files for each quarter (8 total) were in different order, so we reordered all the files.
* Format all the columns to keep the standardization in the data.

This data process was done looking at the data quality issues, such as accuracy and believability.

Finally, it was necessary to integrate these 8 files sources in only one large CSV file. To merge and combine theses large files we used the CMD tool. The final integrated dataset file is “combined-csv-files\_1B.csv”.



### Figure 4. Data Integration with CDM.

The integrated dataset file “combined-csv-files\_1B.csv” includes **3,415,323** rows and 14 columns:

Count\_Index

trip\_id

trip\_start\_time

trip\_start\_date

t\_start\_time

trip\_stop\_time

trip\_stop\_date

t\_stop\_time

trip\_duration\_seconds

from\_station\_id

from\_station\_name

to\_station\_id

to\_station\_name

user\_type

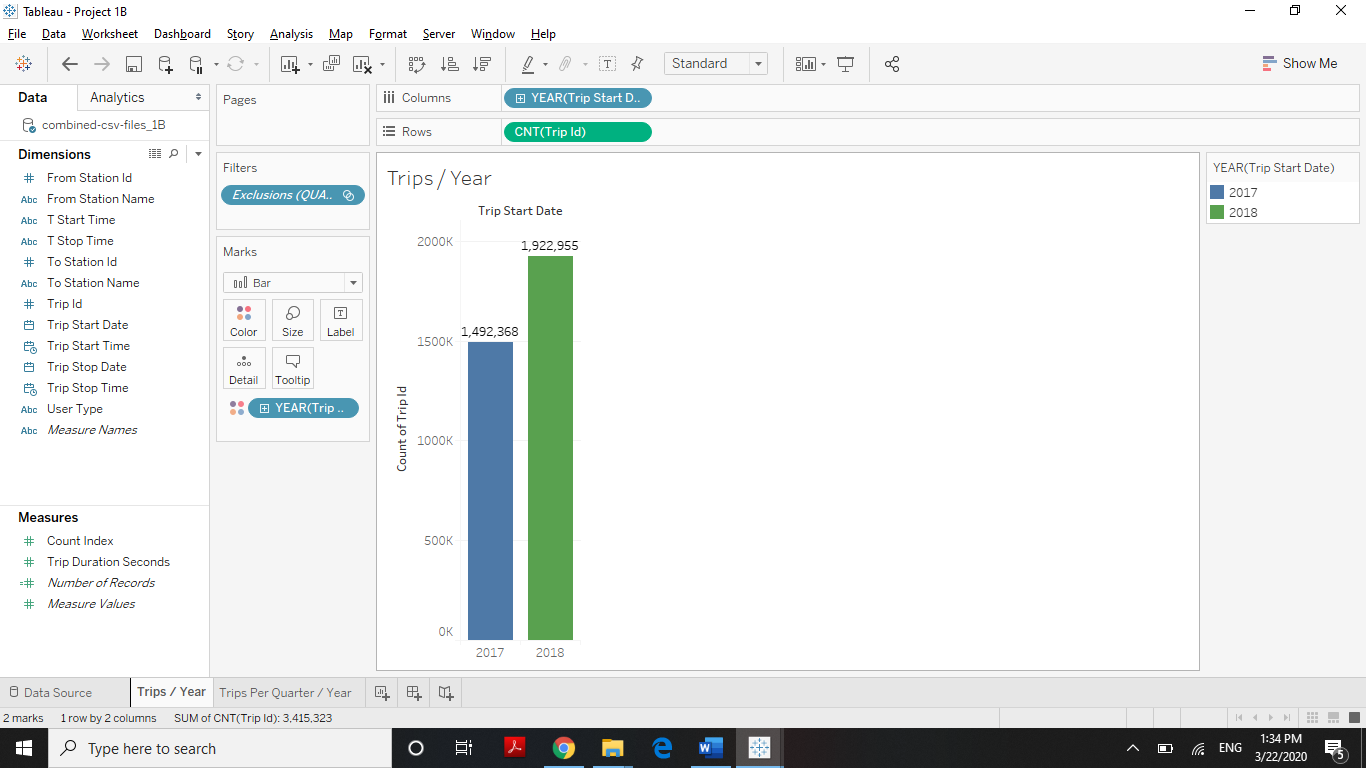
* Data reduction:

There is no smaller subset of data that can produce the same or similar analytical results.

Data processing is one of the longest tasks in data analysis, this phase took about 24 hours to get the final cleaned dataset to make the required analysis.

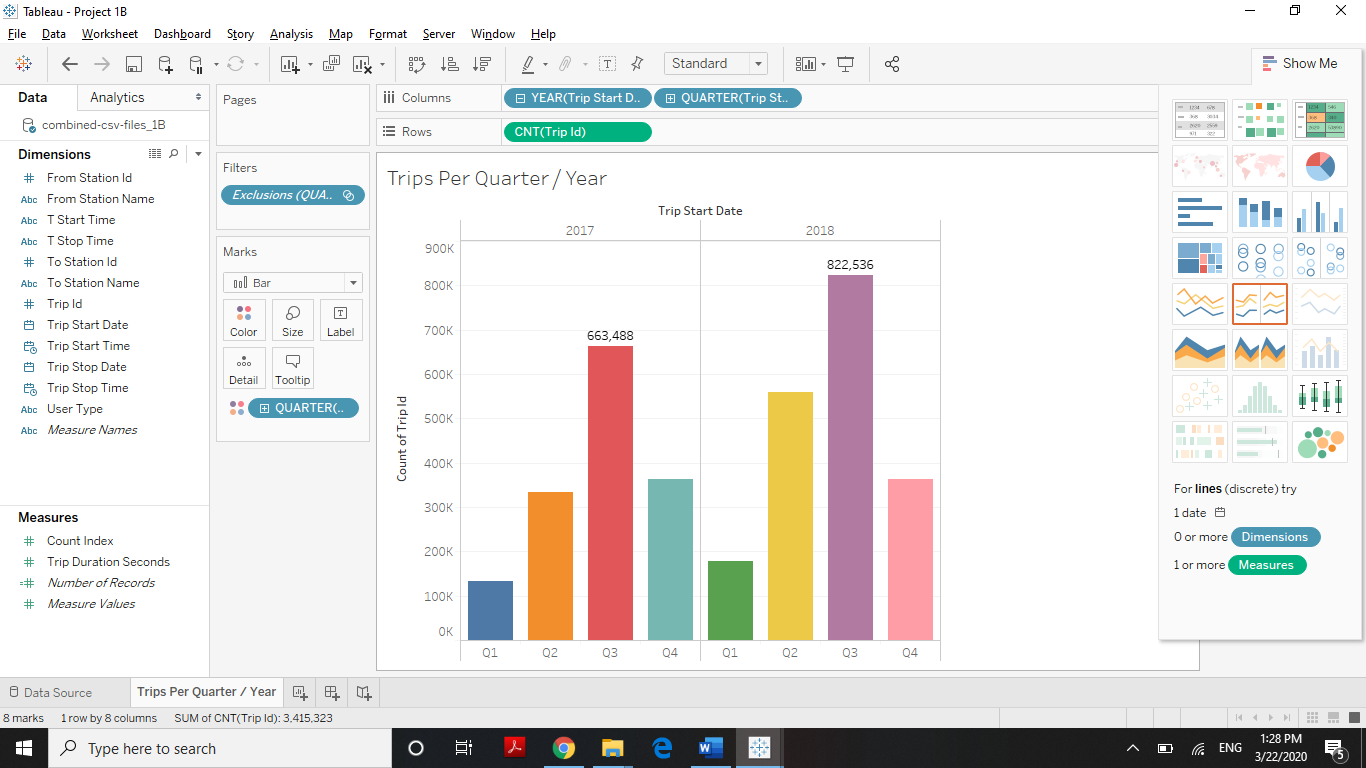
* **Current State of Key Y Outputs**

With the dataset cleaned and preprocessed, we used Tableau to get the following analysis:



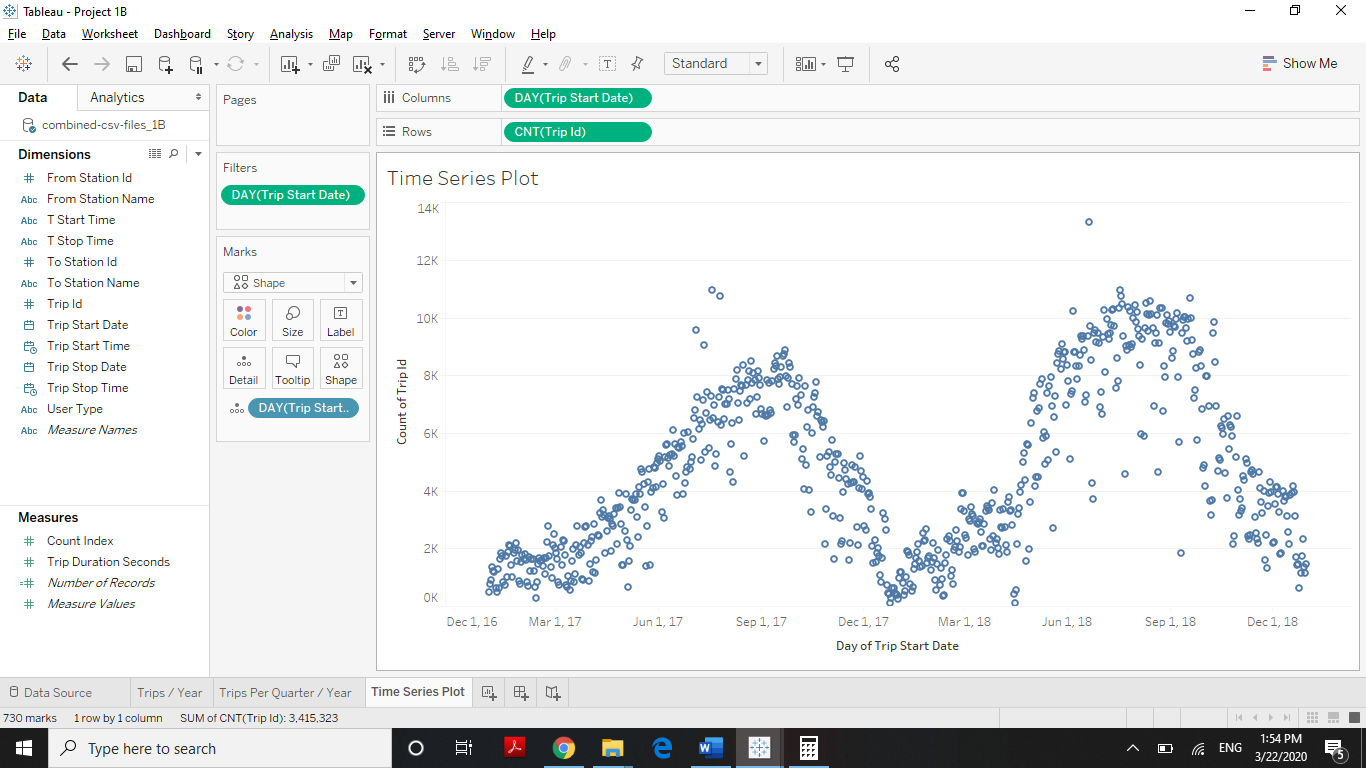
### Figure 5. Trips per Year.

The quantity of trips was in 2017 = 1,492,368 2017 and 2018 = 1,922,955. The increase was 430,587 trips, or 28.85%



### Figure 6. Trips per Quarter / Year.

For both, 2017 and 2018, Quarter 3 was the highest period: 2017\_Q3= 663,488; 2018\_Q3: 822536 trips.



### Figure 7. Time Series Plot.

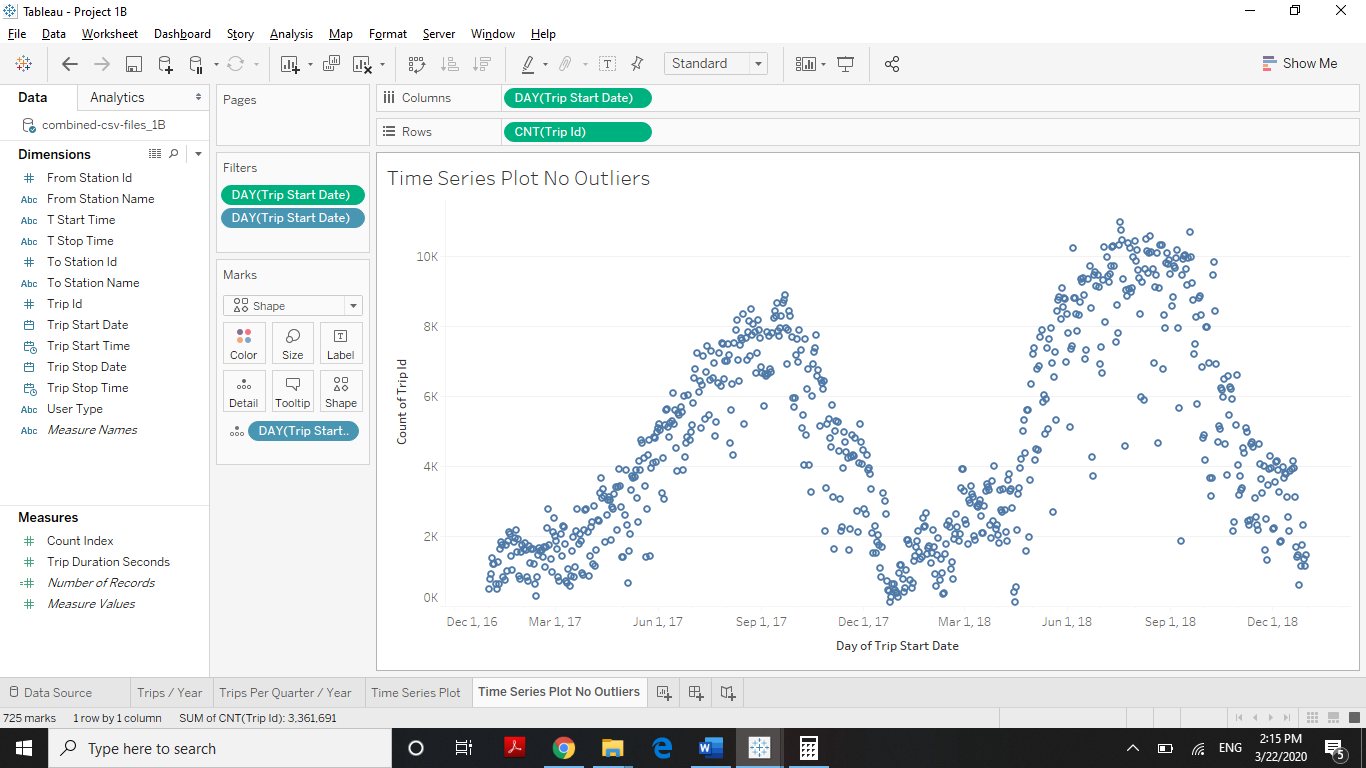
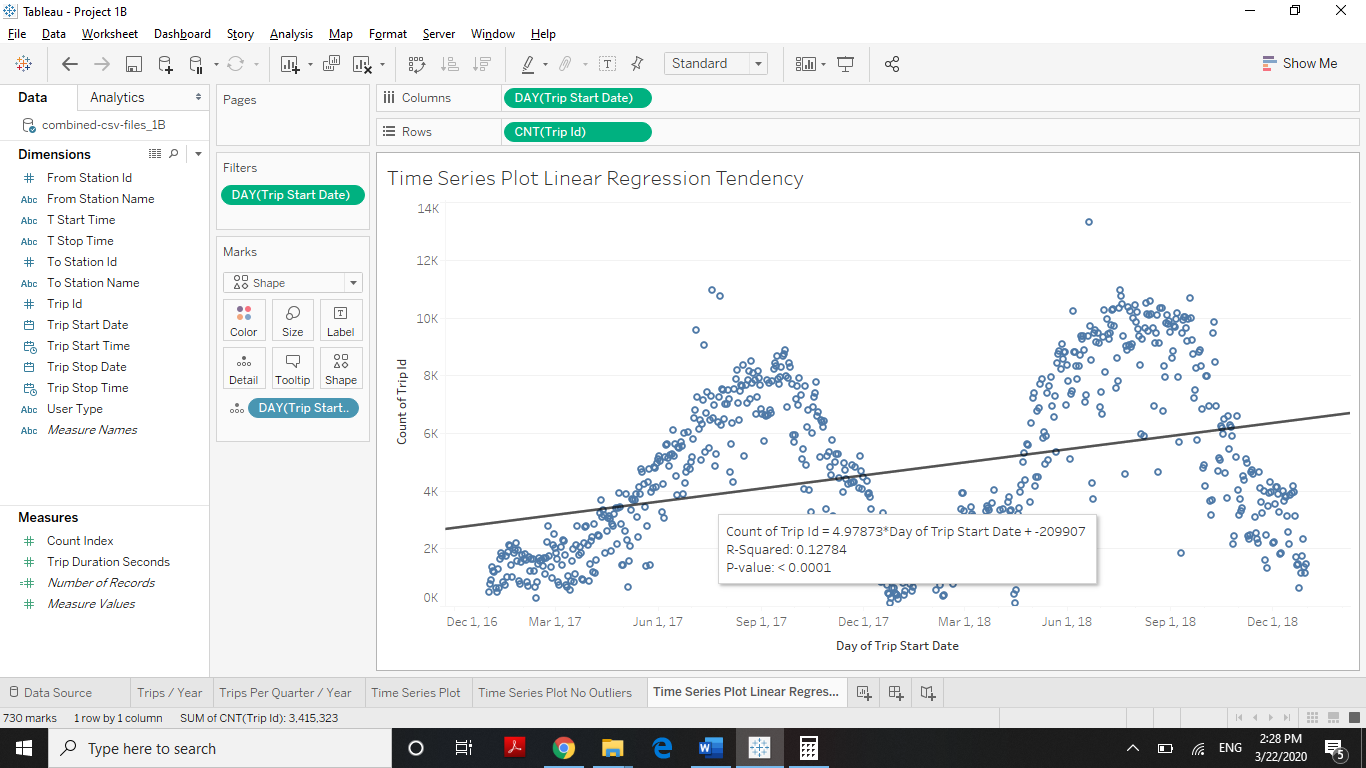


Figure 8. Time Series Plot Not Outliers.

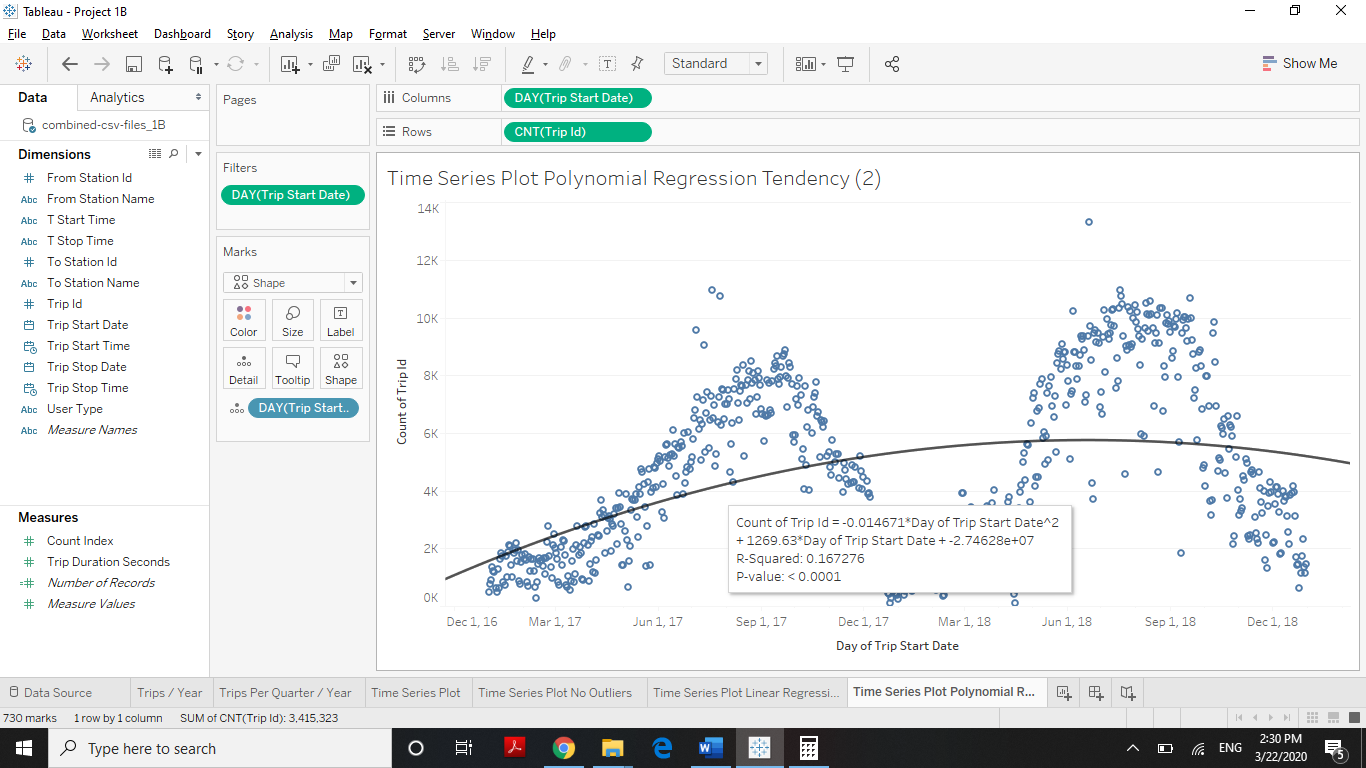
This time series plot for 2017 and 2018 shows a clear seasonalized behavior in the data. This was also showed in Trips per Quarter/Year graphic. Likewise, there are some outliers, however, if we take them out, the plot keeps the same behavior.

* **Distribution/Data Patterns of Key Y Outputs**



### Figure 9. Time Series Plot Linear Regression Tendency.

Although The linear regression model for this data has a Pvalue < 0.005, the R-Squared is 0.12784, so this model explains the data in only 12.78%. Therefore, this model is not significant.

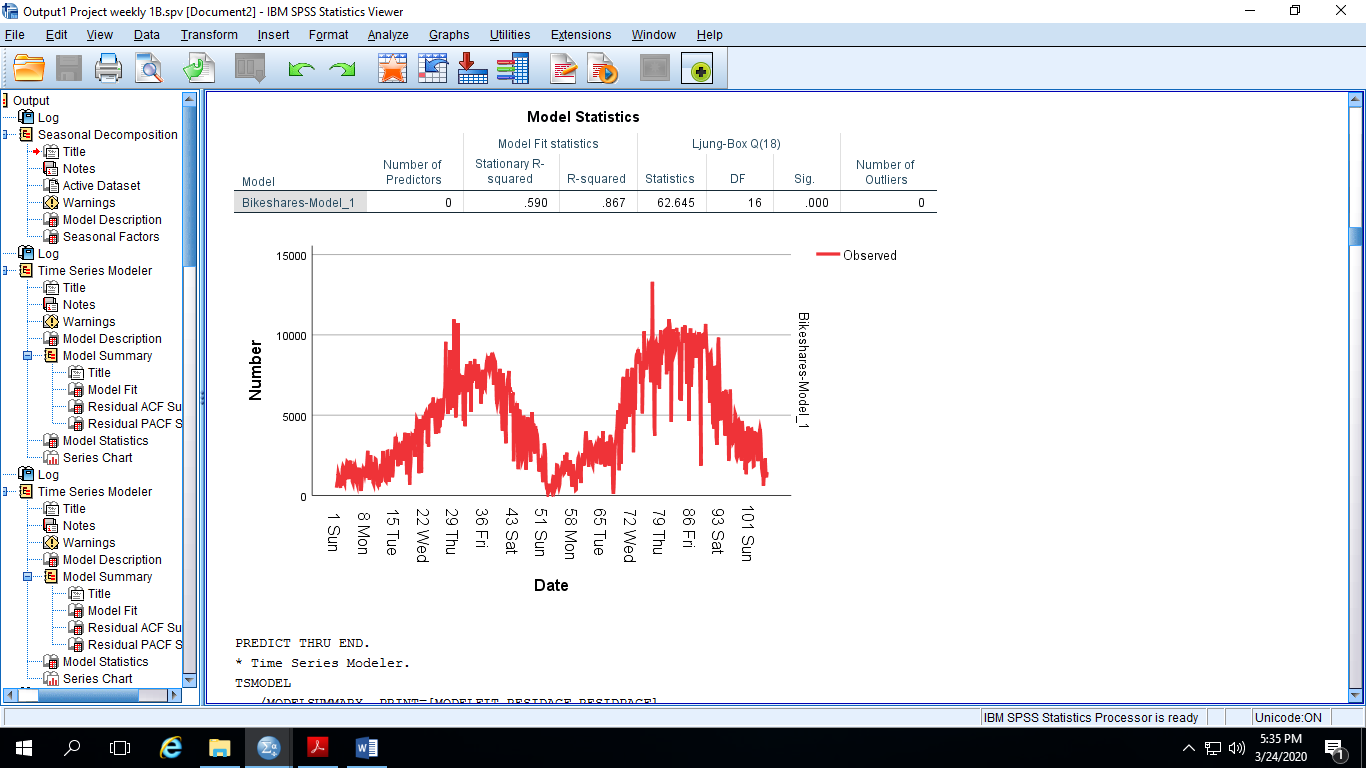


### Figure 10. Time Series Plot Polynomial Regression Tendency.

The polynomial regression model for this data has a Pvalue < 0.005, the R-Squared is 0.167276. This model is better than the LR model, it explains the data in 16.72%. However, this model is not significant enough.

Due to the data distribution, we used SPSS to analyze this dataset. The results are shown as follows:

* Time series plot of the daily count of total rental bikes, Jan 2017 to Dec 2018

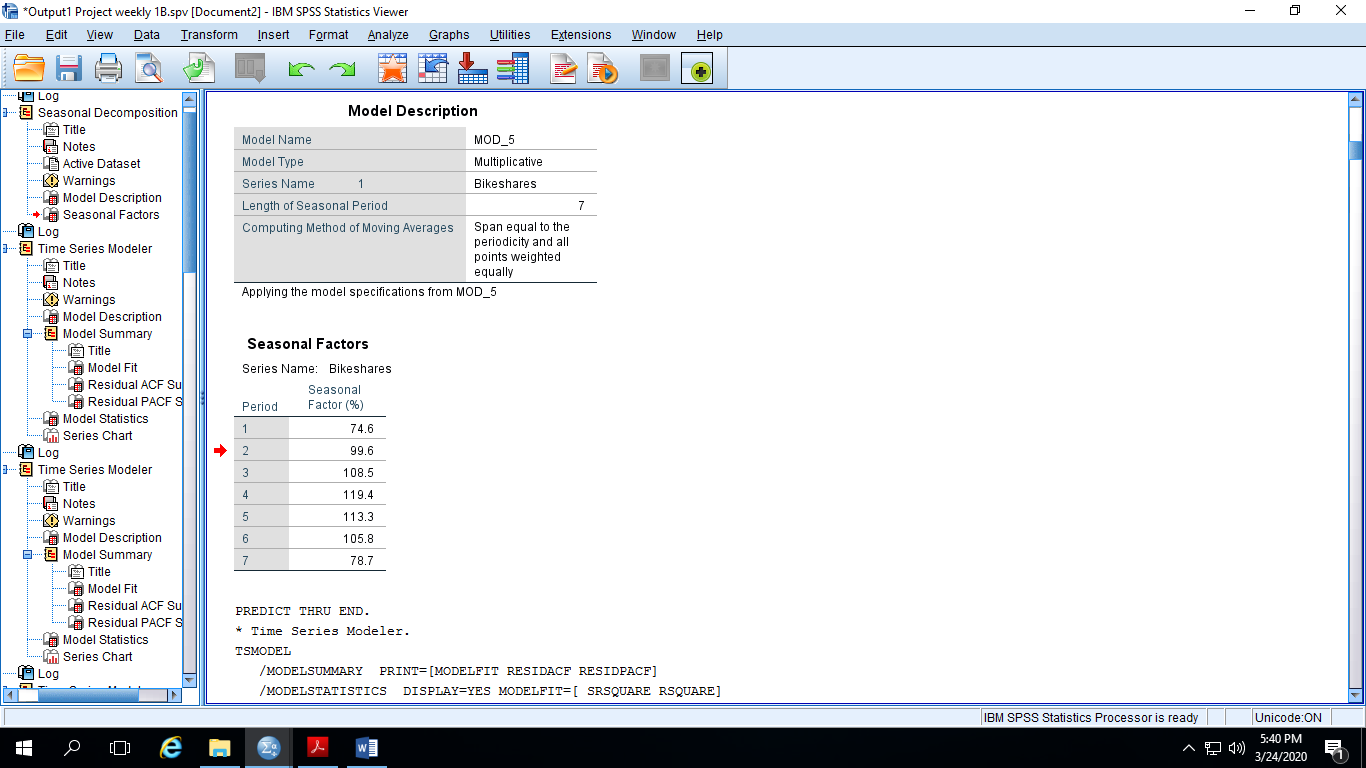
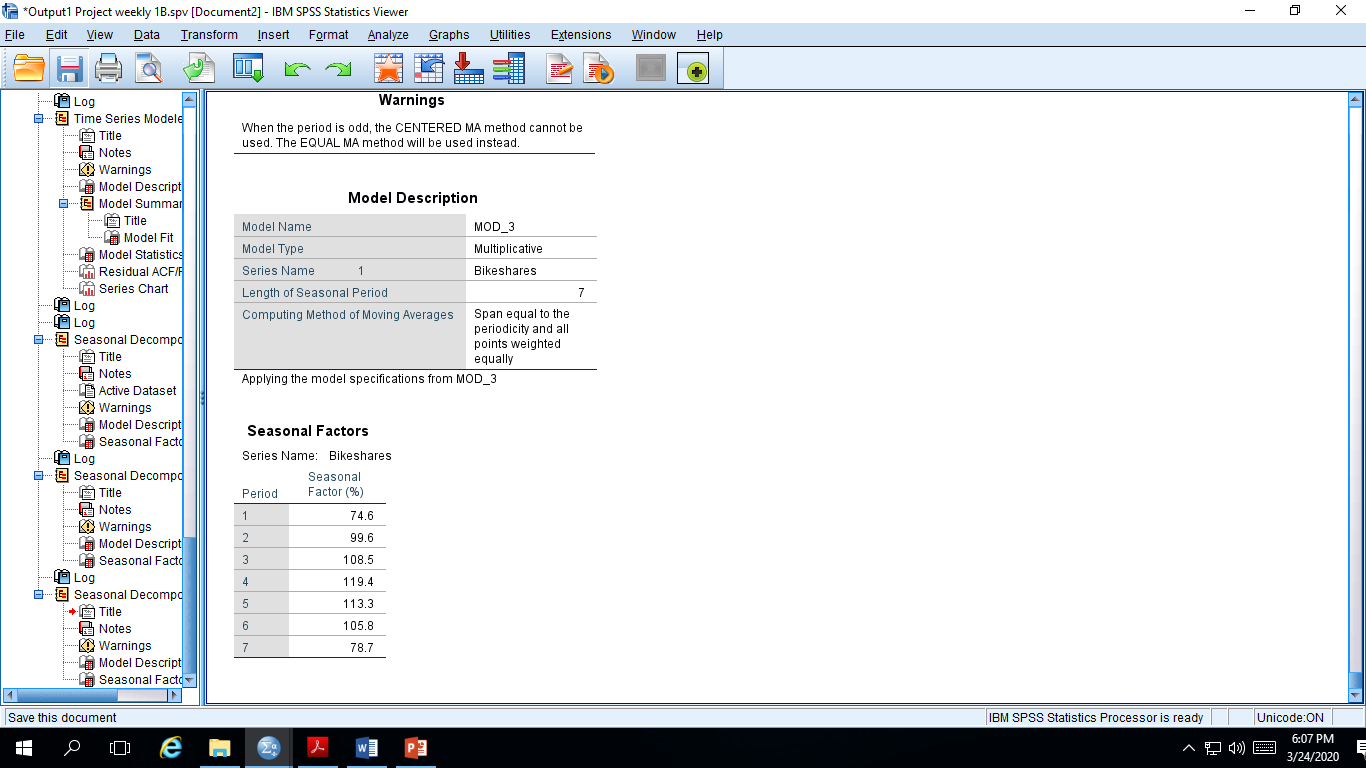


### Figure 11. Time Series Plot from SPSS.

* Time series decomposition:



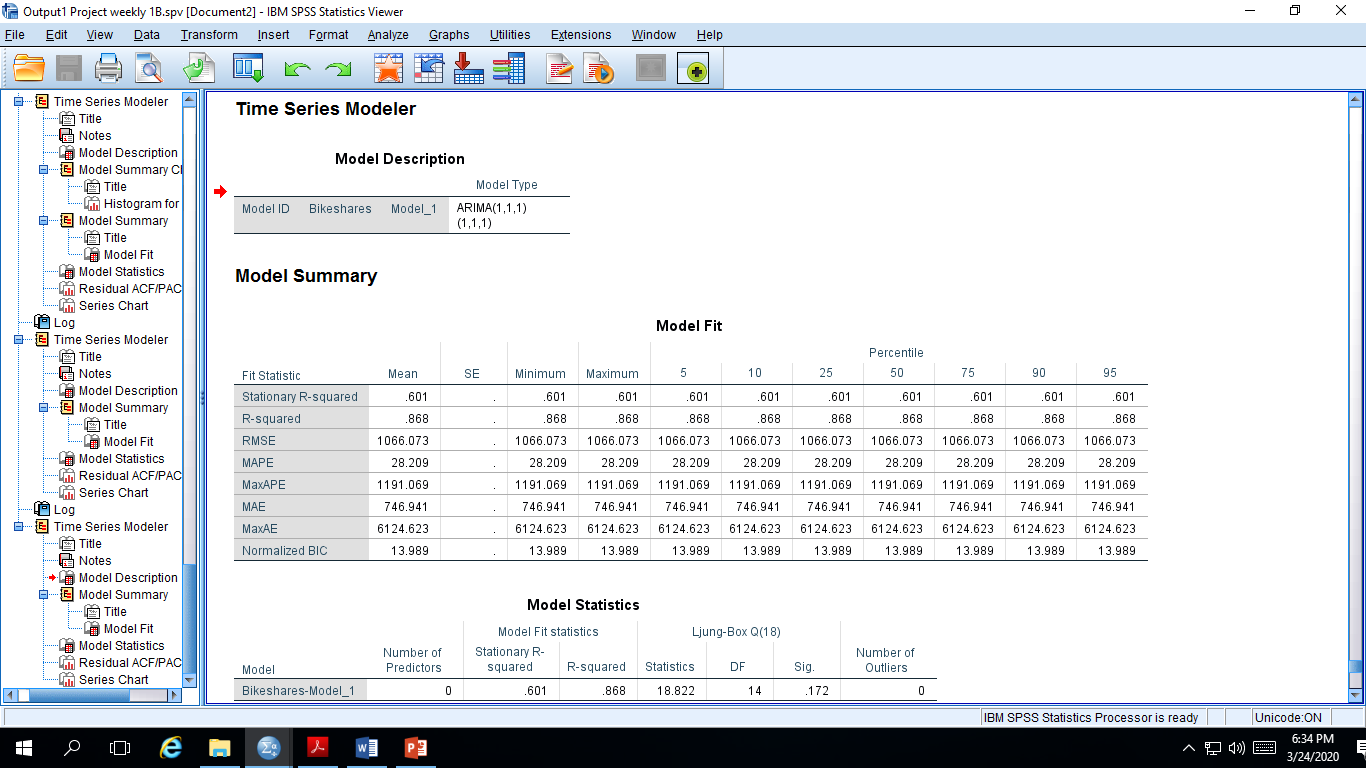
### Figure 12. Time Series Decomposition.



### Figure 13. Time Series Decomposition Model and Factors.

The seasonal indices are the cyclical patterns in the dataset. According with these results, the highest seasonal index is 119.4% of rental volume Wednesdays, and the lowest seasonal index is 74.6% of rental volume on Sundays.

* Description of the final model

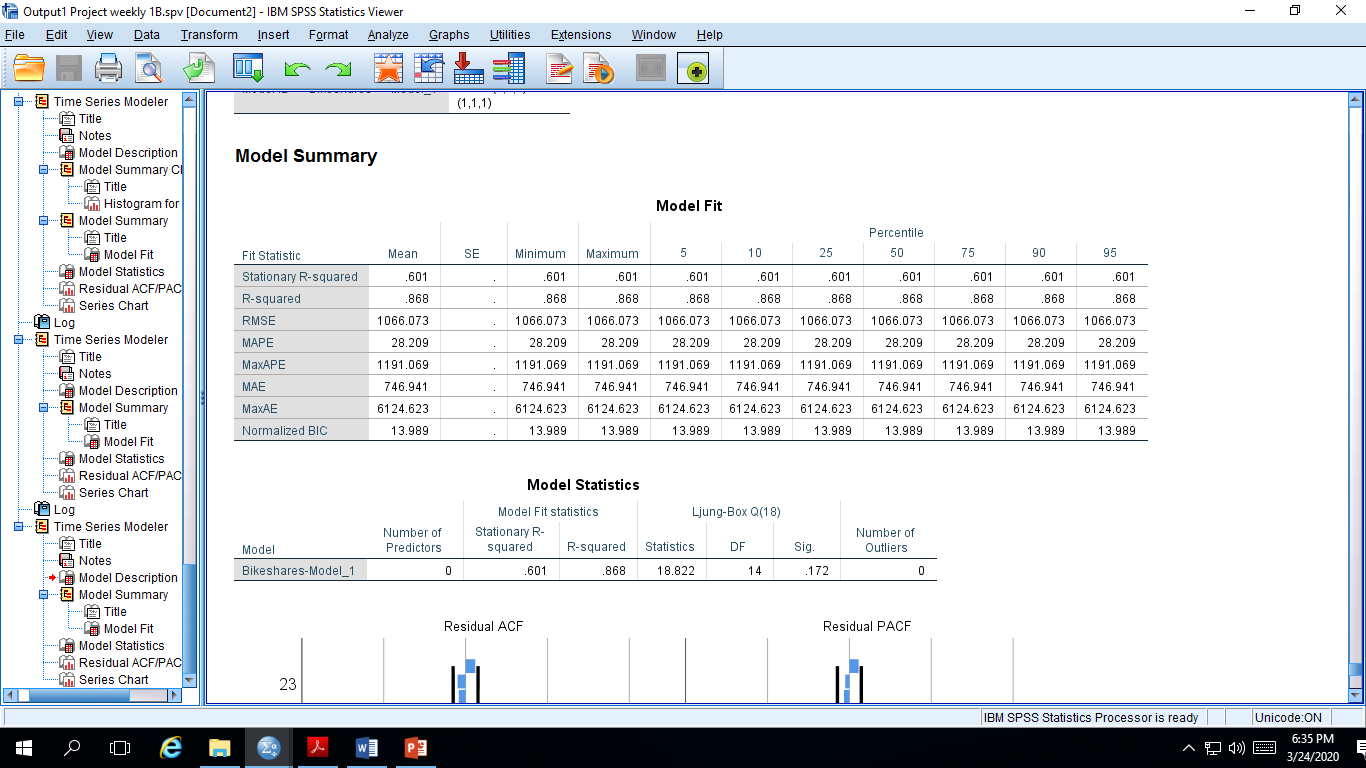


### Figure 14. Final ARIMA Model.

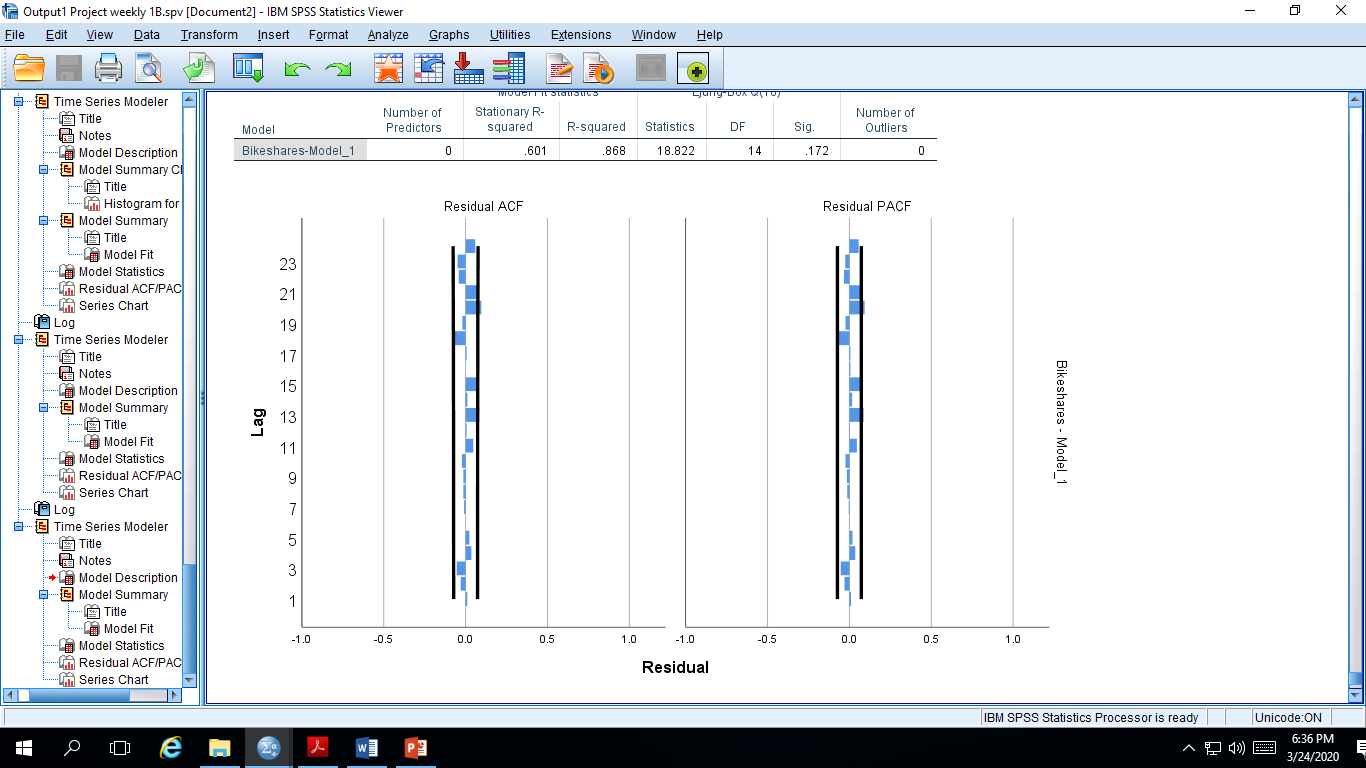
The resulting model is an ARIMA. Its model type is the following:

Non-seasonal part*: p* = 1, *d* = 1, *q* = 1

Seasonal part*: P* = 1, *D* = 1, *Q* = 1



### Figure 15. ARIMA Model Statistics.



### Figure 16. Residual ACF and PACF Plots

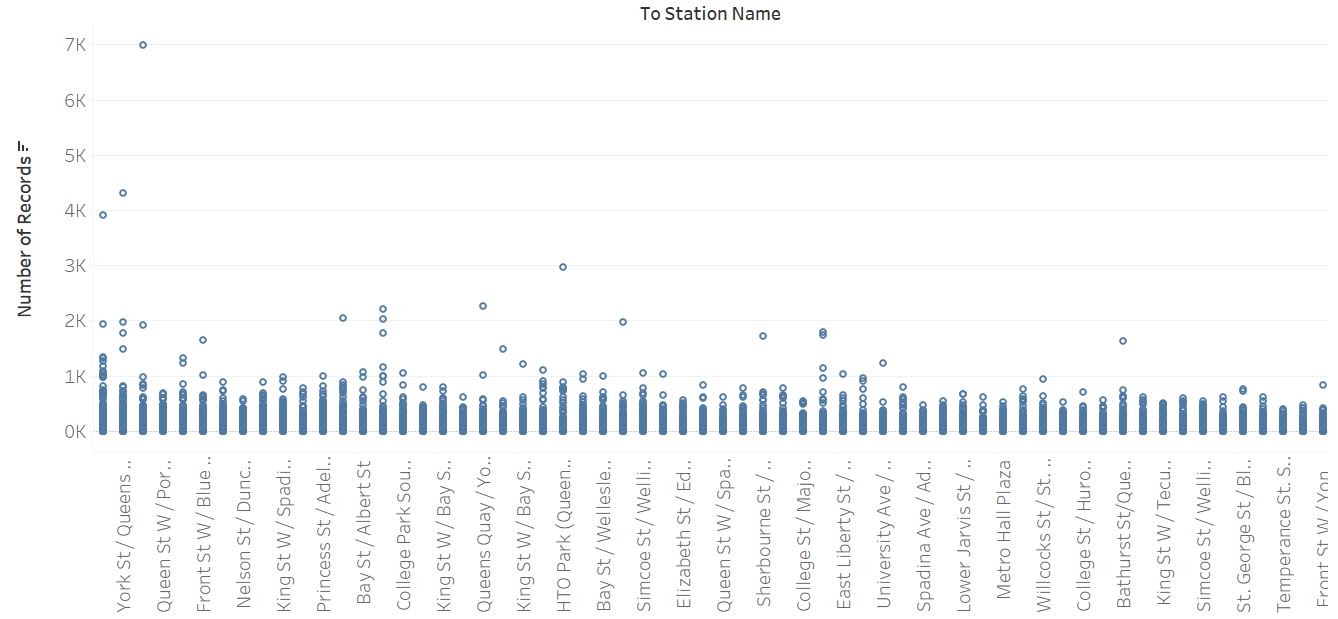
The model R-Squared = 0.869, therefore, the model can explain the data in 86.9%. In addition, the Ljung-Box shows a p-value = 0.172 that is also high. Also, the Residual ACF and PACF plots satisfy the stationarity for the residuals.

In conclusion, the ARIMA model seems to be adequate.

# 3.0 Analysis and Findings: The Analyze Phase

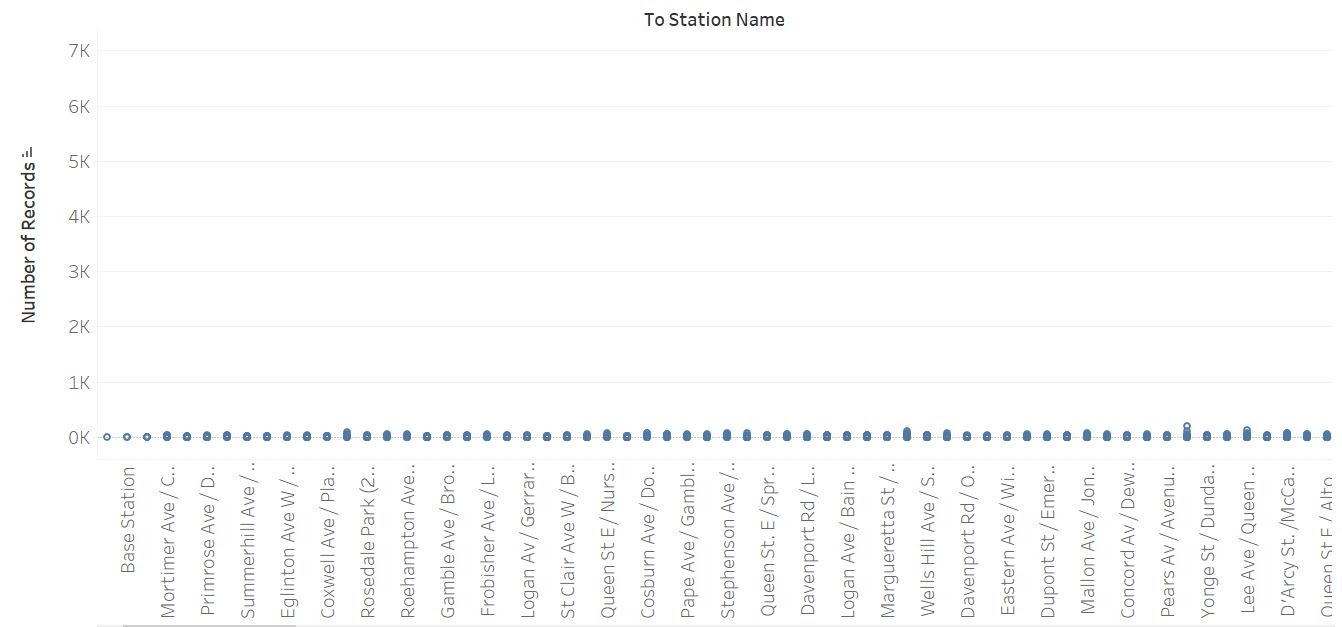
We analyzed the data in Tableau to answer a few key questions that can be used to improve the ridership business. They can be listed as follows:

1. **Most used routes -** Identifying the most used routes is key to finding answers to specific questions like:
2. Why do commuters prefer to use bikes from point A to point B?
3. Can we plot more bikes in these routes to get more business?
4. How to scale the current business model keeping the nearby hotspots in mind?

****

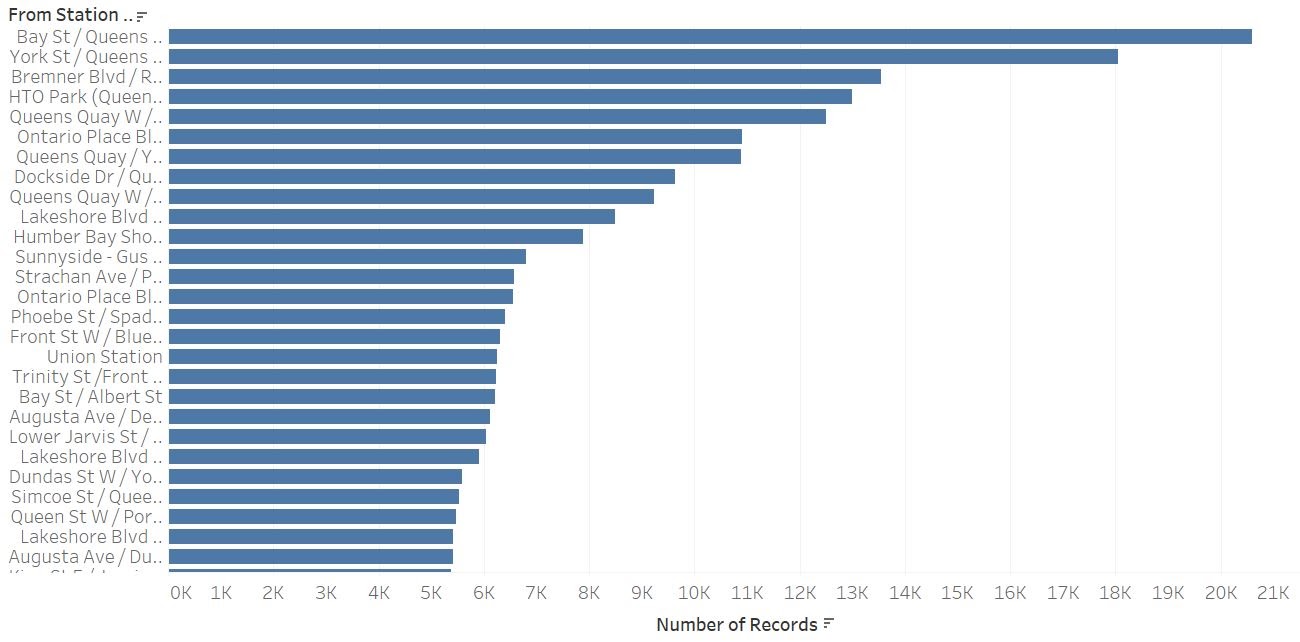
### Figure 17. Most Used Routs.

1. **Least used route -** Determining the least used route may not be a waste of time. The knowledge about the shortcomings of a business are equally important to the answers of questions which can grow your business. Some of the questions can be:
2. Are they near a local transport service?
3. Trip durations in these points are longer than average or shorter?

****

### Figure 18. Least Used Routs.

1. **Targeting casual members -** Converting casual members to Annual members should always be a priority as it has a direct effect on company valuations and revenue.

****

### Figure 19. Maximum Commute To/From Station Wise.

The first part of the analysis shows the most used routes by casual members. The second part, however, is very interesting.In order to find out casual user behavior with respect to time based usage, we create a calculated field using the below mentioned code:

IF [Trip Duration Seconds]<=60 THEN "1 min less than"

ELSEIF [Trip Duration Seconds]>60 AND [Trip Duration Seconds]<=300 THEN "5 mins less than"

ELSEIF [Trip Duration Seconds]>300 AND [Trip Duration Seconds]<=600 THEN "5-10 mins"

ELSEIF [Trip Duration Seconds]>600 AND [Trip Duration Seconds]<=900 THEN "10-15 mins"

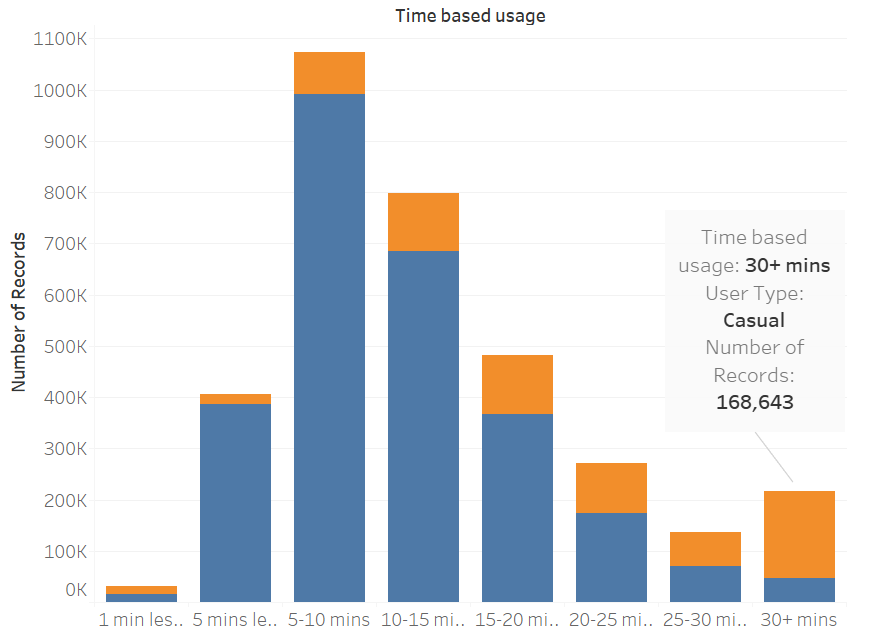
ELSEIF [Trip Duration Seconds]>900 AND [Trip Duration Seconds]<=1200 THEN "15-20 mins"

ELSEIF [Trip Duration Seconds]>1200 AND [Trip Duration Seconds]<=1500 THEN "20-25 mins"

ELSEIF [Trip Duration Seconds]>1500 AND [Trip Duration Seconds]<=1800 THEN "25-30 mins"

ELSE "30+ mins"

END

****

### Figure 20. Time Based Casual Member Usage.

If we look closely, the ratio of casual: annual increases as the duration of trip increases. It’s almost equal in the 25-30 min interval which tells us that most casual users prefer using the service for longer durations. With proper incentives and strategy, we can convert these users to annual members.

# 4.0        Recommendations: The Improve Phase:

          When we want to improve any business, we have to think about the problems in that and    provide the improvements. These are the improvements we consider to make up the bike share business in Toronto:.

1. Reviews of the Customers: After completing the ride we need to make sure that the customer is providing a review to say how the ride was and any changes or improvements they want to suggest. These reviews help us to figure out the problems and implement them.
2. Marketing and Information: The informational aspect of this initiative involves letting people know about the service and how to use it. We need to promote the bike services to more people so that it is accessible to more people. We need to attract the people by providing them different offers and making them know the benefits of these bikes.
3. Identifying Service Gaps: This will help us to install more bikes if needed and operate them according to the time bound. By identifying these service gaps we can improve our services to more customers and see there is no trouble for them.
4. Membership for the customers: We have to see that all the casual members who are just using the bikeshare at times to make them as an annual member. We can attract them by providing exclusive offers which are beneficial for them as well as us.

5.0        Monitoring and Controlling**:**

To take control over the improvisation phase of different factors there are different control plans. According to each improvement the control plan will be different.

1. By finding out the customer reviews we can monitor the efficiency of our bike share ridership. While looking into their suggestions we can implement them on the ridership and attract more customers. Finally, we create a response plan in case there are any uncertain changes. We can have a response plan for this control phase, but it depends on the previous data we analyze.
2. Through the marketing and information, we can analyze how many bikes we need more and control the usage accordingly. By providing different promotional offers and see that it attracts all the customers so that the members using our bikeshare becomes more and it gets expanded.
3. By identifying the service gaps in each trip, we can monitor the time of the travel and see that the bikes are available to all the customers who want to travel in a certain area.

# 6.0 Conclusions

This project is implemented through the DMAIC (define, measure, analyze, improve and control) cycle, that is the control tool for six sigma. From the Toronto bike share ridership, we can see the usage of bikes and the number of members who are using the service. We can remove the following conclusions from the above Analysis:

* In both, 2017 and 2018, the quarter 3 (Q3) is the most profitable period for the company. Likewise, we see considerable growth in demand for the service in quarter 2 (Q2).
* The ARIMA model satisfies 86.9% of the data set and can be used as a significant model.
* The most used routes are located in most of the downtown and its surroundings. These routes are very close to the train stations.
* Casual members of the bike-share use the service mainly when the ride duration is more than 30 mins.

# Work Cited

Matthews, K. (2016, December 13). How to Merge Multiple CSV Files and Combine Them Into One Large CSV File. Retrieved March 15, 2020, from <https://www.live2tech.com/merge-multiple-csv-files-combine-one-large-csv-file/>

Navaneeth Mohan (2019, November 17). User behavior of bikeshare data using unsupervised machine learning, from <https://medium.com/@navaneeth.mohan94/understanding-user-behavior-of-torontos-bike-share-an-exercise-in-unsupervised-machine-learning-8dd858e1cdcb>

JackyWang. (2020, February 4). Toronto Bikeshare Data. Retrieved March 15, 2020, from <https://www.kaggle.com/jackywang529/toronto-bikeshare-data/data>

# Appendix A.

## Statistical Analysis Plan.

View the following pages.

**Statistical Analysis Plan**

**For**

**Business Process Analysis for Bike Share Toronto Ridership**

DSCI 5260 Business Process Analytics

University of North Texas

**Version:** Project No. 1

**Author**:  Samuel Castilla

Satvika Marrapu

Suhail Bari

**Date:** 01-April-2020

**Table of Content**

[1 INTRODUCTION TO YOUR PROJECT WITH BASIC BACKGROUND 23](#_Toc36564228)

[2 DATA SOURCE 23](#_Toc36564229)

[3 ANALYSIS OBJECTIVES 24](#_Toc36564230)

[4 ANALYSIS SETS/ POPULATIONS/SUBGROUPS 24](#_Toc36564231)

[5 ENDPOINTS AND COVARIATES 24](#_Toc36564232)

[6 HANDLING OF MISSING VALUES, OUTLIERS AND OTHER DATA CONVENTIONS 24](#_Toc36564233)

[7 STATISTICAL METHODOLOGY 25](#_Toc36564234)

[7.1 STATISTICAL PROCEDURES 25](#_Toc36564235)

[7.2 MEASURES TO ADJUST FOR MULTIPLICITY, CONFOUNDS, HETEROGENEITY, ETC.\_…………………………………………………………….26](#_Toc36564236)

[8 SENSITIVITY ANALYSES 26](#_Toc36564237)

[9 RATIONALE FOR ANY DEVIATION FROM PRE-SPECIFIED ANALYSIS PLAN PERFORMED 26](#_Toc36564238)

[10 PLANS TO ENSURE QUALITY AND ETHICS 27](#_Toc36564239)

[11 PROGRAMMING PLANS (USE OF PYTHON, R, etc.) 27](#_Toc36564240)

[12 REFERENCES USED IN THE COURSE OF YOUR ANALYSES CAN BE FOUND IN APPENDICES UNDER "APPENDIX B" 28](#_Toc36564241)

[13 APPENDICES 28](#_Toc36564242)

[APPENDIX A 29](#_Toc36564243)

[OUR FIRM'S ETHICAL CODE 30](#_Toc36564244)

[APPENDIX B 32](#_Toc36564245)

[REFERENCES 33](#_Toc36564246)

**CONTENTS**

1. INTRODUCTION TO YOUR PROJECT WITH BASIC BACKGROUND

Bikes are the most convenient mode of transport that dates back to the early 19th century. In the modern day, with the inception of so many new media of transport, they still play a vital role in our dynamic society. One of the businesses that is taking the spotlight in today’s market is related to bike rentals. Bike rentals have changed over the years from the traditional in-person payment and pick-up to a totally automatic rent and return service. These fully automated systems offer the user the convenience of renting a bike in from any of the pickup spots and return them for any duration of time, preferred by the user.

This project is about analyzing Toronto’s bike share ridership data from 2017 and 2018. Based on the obtained results from this analysis, the project includes the study of possible business aspects such as most convenient stops (as per user data), peak days of the week, how to convert users to annual plan and decide pricing based on duration of trip.

1. DATA SOURCE

The data set provides information about the bike share ridership. This dataset contains information from 2017 and 2018 and is divided in quartiles for each year. It has 3,415,324 records in total. The Bike Share dataset includes anonymized trip data, including Trip start date and time, Trip end date and time, Trip duration, Trip start station, Trip end station, and User type.

According to Kaggle, “this dataset is from Toronto Parking Authority, published on https://open.toronto.ca/dataset/bike-share-toronto-ridership-data/. The data is licensed under: Open Government License – Toronto”.

The source of this secondary dataset is <https://www.kaggle.com/jackywang529/toronto-bikeshare-data/data>

Data dictionary

trip\_id -- A unique id created for each trip

trip\_start \_time -- The start date and time of the trip

trip\_stop\_time -- The end date and time of the trip

trip\_duration\_seconds -- The time duration of the trip in seconds

from\_station\_id -- A unique id for each station

to\_station\_id -- A unique id for each station

from\_station\_name -- Name of the start station

to\_station\_name -- Name of the end station

user\_type -- Types of users

1. ANALYSIS OBJECTIVES

The overall scientific objectives of the analysis are the following:

* Determine the correlations between variables within the dataset.
* Draw the time series plot of the data to see its behavior in the analyzed period.
* Make the time series decomposition to determine relevant factors such as seasonal indices and trend and cyclical components.
* Determine the model that best explain the dataset.
* Based on the identified model, study of some possible business processes such as business planning, shipping, and warranty and claims processing.

1. ANALYSIS SETS/ POPULATIONS/SUBGROUPS

Since the dataset does not include many variables, the project includes all of them. However, data will be divided according to user type (member/casual) and we would study the dataset based on each quartile due to the possible seasonal behavior.

1. ENDPOINTS AND COVARIATES

The project does not include endpoints nor covariates.

1. HANDLING OF MISSING VALUES, OUTLIERS AND OTHER DATA CONVENTIONS

To handle this data cleaning, we will consult the following references:

Newton, R. R., & Rudestam, K. E. (2013). Your statistical consultant: Answers to your data analysis questions. Thousand Oaks: SAGE Publications.

Konasani, V. R., & Kadre, S. (2015). Practical business analytics using SAS: A hands-on guide.

In addition, to handle missing values the project will use the “most probable value” approach. Therefore, we will estimate these missing values using regression, decision tree, etc.

For example, using SAS Enterprise miner, the SAS Impute button is used to impute missing values. The Tree surrogate option gives the best outcome when compared with Andrews waive, Distribution, Tukey’s Bi-Weight, Mean and Huber. The Tree surrogate option enables SAS to first understand existing values and build a fitted predictive model to predict missing values.

About noisy data, the project includes the following possibilities:

* Binning: sort and adjust the value based on those of its neighbors (mean, median, boundary).
* Regression: Use predicted rather than actual values.
* Outlier analysis: Identify and exclude “odd” records.

To analyze the variables, we will use box plot to check for abnormal values. Also, SAS enterprise miner Filter option brings a filter for all outliers. Likewise, we will plot the dataset to find any noisy data.

1. STATISTICAL METHODOLOGY

A standard statistical procedure that will concur a relationship between two quarters of the same year and identical quarters from two different years i.e 2017 and 2018. The Null hypothesis theory will be tested to determine any relationship based on variables between the data sets using cluster analysis.

For a particular quarter, we plan to analyze the properties of the distribution and find out any outliers and the form theories as to why they became such outliers.

* 1. STATISTICAL PROCEDURES

Depending on the variables, we plan to use correlation, regression analysis and ANOVA between two quarters of data. The method or procedure can be replicated from the SPSS tutorials website referenced below:

* Degree of Relationship:

To determine the relationship and correlation between the quantitative variables included in the dataset, this project uses the statistical test named Multiple Regression.

* Significance of Group Differences:

To determine the causal relationship between the dependent and the independent variables in the project, we may use the t test, or the analysis of variance (ANOVA) test.

* Prediction of Group Membership:

To identify the independent variables that best predict the dependent variable the project may use the discriminant analysis

* 1. MEASURES TO ADJUST FOR MULTIPLICITY, CONFOUNDS, HETEROGENEITY, ETC.

As of now, this point is invalid. If the data gives out these symptoms, we will adhere to standard statistical research practices.

For Multiplicity, we can refer to multiplicity issues in clinical trials: the what, why, when and hows taken from an article from the International Journal of Epidemiology. A methodological guidance review from BMC Med Res Methodol can be used to study the clinical aspects of heterogeneity.

The following are the references:

Guowei Li,1,2,\* Monica Taljaard,3,4Edwin R. Van den Heuvel,5,6Mitchell AH. Levine,1,2,7Deborah J. Cook,1,2,7George A. Wells,3,8Philip J. Devereaux1,7,9and Lehana Thabane1,2,9 - An introduction to multiplicity issues in clinical trials: the what, why, when and how.

https://www.researchgate.net/publication/311947626\_An\_introduction\_to\_multiplicity\_issues\_in\_clinical\_trials\_the\_what\_why\_when\_and\_how

BMC Med Res Methodol. 2012; 12: 111 - Investigating clinical heterogeneity in systematic reviews: a methodologic review of guidance in the literature

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3564789/

1. SENSITIVITY ANALYSES

In this particular project, the data is secondary and no new/uncertain inputs have to be given. Hence, we don’t need to perform a sensitivity analysis.

1. RATIONALE FOR ANY DEVIATION FROM PRE-SPECIFIED ANALYSIS PLAN PERFORMED

The data will not be changed except for additions of a few columns for real world calculation purposes.

1. PLANS TO ENSURE QUALITY AND ETHICS

Please refer to Appendix A “Our Firm’s Ethical Code”.

1. PROGRAMMING PLANS (USE OF PYTHON, R, etc.)

To analyze and process data, we will use the following coding:

Using Python:

*#visualising numeric data using a histogram*

**%**matplotlib inline

**import** matplotlib.pyplot **as** plt

data.hist(bins**=**20, figsize**=**(20,15))

**from** pandas.plotting **import** scatter\_matrix

attributes **=** ["Variable name", "Variable name"]

scatter\_matrix(data[attributes], figsize**=**(12,8))

*#plotting the data (for visualization purposes only)*

**%**matplotlib inline

**import** matplotlib

**import** matplotlib.pyplot **as** plt

plt.rcParams['axes.labelsize'] **=** 14

plt.rcParams['xtick.labelsize'] **=** 12

plt.rcParams['ytick.labelsize'] **=** 12

plt.plot(X, y, "b.")

plt.xlabel("$x\_1$", fontsize**=**18)

plt.ylabel("$y$", rotation**=**0, fontsize**=**18)

plt.axis([0, 1, 4, 12])

plt.show()

*#imputing missing values*

*#scaling numeric data*

**import** numpy **as** np

**from** sklearn.preprocessing **import** StandardScaler

**from** sklearn.preprocessing **import** Imputer

imputer**=** Imputer(strategy**=**"median")

scaler**=**StandardScaler()

col\_names**=**list(df\_num)

num\_col**=**np.array(df\_num)

num\_col\_imp**=**imputer.fit\_transform(num\_col)

num\_col\_scaled**=**scaler.fit\_transform(num\_col\_imp)

num\_col\_scaled

df\_num\_scaled**=**pd.DataFrame(num\_col\_scaled, columns**=**col\_names)

df\_num\_scaled.head()

*#fitting linear regression by solving the normal equation using simple numpy*

*#.dot is a numpy method for matrix multiplications*

X\_b **=** np.c\_[np.ones((N, 1)), X] *# add x0 = 1 to each instance*

theta\_best **=** np.linalg.inv(X\_b.T.dot(X\_b)).dot(X\_b.T).dot(y)

theta\_best

*#making predictions*

X\_new **=** np.array([[0], [1]])

X\_new\_b **=** np.c\_[np.ones((2, 1)), X\_new] *# add x0 = 1 to each instance*

y\_predict **=** X\_new\_b.dot(theta\_best)

y\_predict

1. REFERENCES USED IN THE COURSE OF YOUR ANALYSES CAN BE FOUND IN APPENDICES UNDER "APPENDIX B"
2. APPENDICES

**APPENDIX A**

**OUR FIRM'S ETHICAL CODE**

We pledge in writing to abide by the American Statistical Association's (ASA) and INFORMS' Codes of Ethics. Our adherence to these Codes signifies voluntary assumption of self-discipline. As the professional associations for our firm in the United States, the ASA and INFORMS requires adherence to their Codes of Ethics as a condition of membership. The standards of conduct set forth in these Codes provide basic principles in the ethical practice of data analysis consulting. The purpose of these Codes is to help us maintain our professionalism and adhere to high ethical standards in the conduct of providing services to clients and in our dealings with our colleagues and the public. Our individual judgment requires we apply these principles. We are liable to disciplinary action under the ASA's and INFORMS' Rules of Procedure for Enforcement of this Code if our conduct is found by the ASA's or INFORMS' respective Ethics Committees to be in violation of their respective Codes or to bring discredit to the profession or to ASA and INFORMS .

**Our Commitment to Our Clients**

1. We will serve our clients with integrity, competence, independence, objectivity, and professionalism.
2. We will mutually establish with our clients realistic expectations of the benefits and results of our services.
3. We will only accept assignments for which we possess the requisite experience and competence to perform and will only assign staff or engage colleagues with the knowledge and expertise needed to serve our clients effectively.
4. Before accepting any engagement, we will ensure that we have worked with our clients to establish a mutual understanding of the objectives, scope, work plan, and fee arrangements.
5. We will treat appropriately all confidential client information that is not public knowledge, take reasonable steps to prevent it from access by unauthorized people, and will not take advantage of proprietary or privileged information, either for use by ourselves, the client's firm, or another client, without the client's permission.
6. We will avoid conflicts of interest or the appearance of such and will immediately disclose to the client circumstances or interests that we believe may influence my judgment or objectivity.
7. We will offer to withdraw from a consulting assignment when we believe my objectivity or integrity may be impaired.
8. We will refrain from inviting an employee of an active or inactive client to consider alternative employment without prior discussion with the client. Our Commitment to Fiscal Integrity
9. We will agree in advance with a client on the basis for fees and expenses and will charge fees that are reasonable and commensurate with the services delivered and the responsibility accepted.
10. We will not accept commissions, remuneration, or other benefits from a third party in connection with the recommendations to a client without that client's prior knowledge and consent, and will disclose in advance any financial interests in goods or services that form part of such recommendations. Our Commitment to the Public and the Profession
11. If within the scope of my engagement, we will report to appropriate authorities within or external to the client organization any occurrences of malfeasance, dangerous behavior, or illegal activities.
12. We will respect the rights of consulting colleagues and consulting firms and will not use their proprietary information or methodologies without permission.
13. We will represent the profession with integrity and professionalism in my relations with our clients, colleagues, and the general public.
14. We will not advertise our services in a deceptive manner nor misrepresent or denigrate individual consulting practitioners, consulting firms, or the consulting profession.
15. If we perceive a violation of the Code, we will report it to the APA and INFORMS and will promote adherence to the Code

by other member consultants working on our behalf.

Satvika Marrapu

Suhail Bari

Samuel Castilla

**APPENDIX B**

**REFERENCES**

Newton, R. R., & Rudestam, K. E. (2013). *Your statistical consultant: answers to your data analysis questions*. Thousand Oaks, CA: SAGE.

Konasani, V. R., & Kadre, S. (2015). *Practical business analytics using Sas: a hands-on guide*. New York: Apress.

Li, Guowei & Taljaard, Monica & Van den Heuvel, Edwin & Mitchell, Alexandra & Cook, Deborah & Wells, George & Devereaux, Philip & Thabane, Lehana. (2016). *An introduction to multiplicity issues in clinical trials* ... (n.d.). Retrieved from https://www.researchgate.net/publication/311947626\_An\_introduction\_to\_multiplicity\_issues\_in\_clinical\_trials\_the\_what\_why\_when\_and\_how

Gagnier, J. J., Moher, D., Boon, H., Beyene, J., & Bombardier, C. (2012, July 30). Investigating clinical heterogeneity in systematic reviews: a methodologic review of guidance in the literature. Retrieved March 15, 2020 from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3564789/.

An introduction to multiplicity issues in clinical trials ... (n.d.). Retrieved March 15, 2020, from https://www.researchgate.net/publication/311947626\_An\_introduction\_to\_multiplicity\_issues\_in\_clinical\_trials\_the\_what\_why\_when\_and\_how