UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

3250-Foundations of Data Science

Data Science Fundamentals Certificate

Term Project - TTC Ridership

Group H Members:
Ashokkumar Mistry

Diem Anh Nguyen

Md Mominur Rahaman

Muhammad Raza

Table of Contents

Executive Summary

This report is as part of our Term project for Group H in the course 3250 Foundations of Data Science. In this report, we tried to showcase our learnings from this course. This report elaborates the process that we followed in finding and selecting the dataset from open data source(website) for TTC, transforming and cleaning data, conducting data analysis and reporting key statistical features and insights of the data. We will also focus on the linear regression and ARIMA model. At the end, we will derive an opportunity where TTC can consider in near future.

In the first section, we discuss our objective and covered the overview of our dataset. This part covers the reason why we choose the TTC Ridership data for analysis. We explored multiple datasets and decided to analyze TTC dataset as it is an essential part of daily life in the GTA. In addition, you can get idea on data format, attributes and glimpse of data set with its quality.

Within the Data Preparation section, it describes how we organize the data, how it was cleaned up, the way we handled missing data, transformation and data storing procedure in SQLite. Data in either CSV/Excel format was collected and then we used Anaconda, Jupyter Notebook, Python's Pandas, Matplotlib, Numpy Libraries, sqlalchemy package, sklearn and statsmodel etc. to do our analysis. In our approach-data was read from the source file once in first run and stored in SQLite table after processing. Afterwards, data was retrieved from the SQLite table and there was no need to process in consequent runs. It allowed us to deal with big data efficiently, improve code performance when there are very large datasets to handle. We applied our knowledge and skill sets in data processing and cleaned up the data to use in the analysis phase.

The Data Analysis component include our analysis on the data trend, linear regression, and ARIMA modeling. We analyzed the trend per attribute, outlined the summary of findings from the datasets and figured out the autocorrelation of each attribute. Finally, we used ARIMA model and linear regression to test our datasets. Based on the availability of other datasets, we correlated what TTC can do to enhance the projected traffic.

Finally, we highlighted the challenges and drew a conclusion.

1.0 Introduction

This report represents the analysis of open data source (Ridership on Toronto Transit Commission(TTC), found in City of Toronto Open Data website) as part of Term Project for 3250-Foundation of Data Science. In data analysis, we applied our knowledge that we learn from this course and tried to represent our readiness on analyzing any open data. We choose the open data which has the TTC Ridership information either in excel or CSV or JSON format. It holds the segregation of ridership information measured for TTC. Our Objectives are as follows.

- Analyze trend for overall ridership and revenue
- Forecast on ridership and TTC revenue
- Trend analysis based on passenger type
- Trend analysis based on Fare Media
- Trend Analysis based on types of vehicles
- Trend analysis on Weekday vs Weekend
- Finding Autocorrelation on each attribute
- Using ARIMA model and vet the model with test dataset

2.0 Report Overview

This section of the report explores the avenue we pursued in selecting our data set, displays the overview of the data as well as highlight the data quality.

2.1 Choosing Data Set

We took several avenues to find the right dataset for our project. First, we looked at the free datasets available through the project list. The team realized that most of the datasets are from other countries which we felt were not relevant to us. We also looked at the option of doing an analysis for a company that one of us work for. Unfortunately, due to confidentiality issues, this was not a viable option. Next, we reached out to a corporate contact from OREA, an organization that oversees Ontario Real Estates. As noted before, we wanted to use a dataset that was geographically relevant and did not contain personal confidential information. We ran into another road block as most of the published data have already been analyzed and there was insufficient raw data for us to analyze for this project.

Finally, each member of our team took to further research on the internet. Through Google research, we found several other datasets but one fits the bill. The TTC Ridership dataset had a sufficient number of years of data and features that will allow us to further analyze our dataset for this project.

The dataset was found on the City of Toronto Open Data website, which has the TTC Ridership Analysis in excel/csv/JSON format. It measures the first point of payment by ridership when boarding at the start of a journey using the TTC.

The dataset was downloaded as an Excel spreadsheet. We then imported it to the Panda's Data frame using read_excel method.

2.2 Overview of our Data Set

The chart below shows our original data set. Although it may seem well laid out and clean, it included NaN values, asterisk marks in the column header, calculated rows that are in place to show subtotals and columns with data we do not need. Unfortunately, they add clutter and is not a perfect place to start our analysis. In the following sections, you will find how we addressed these issues.

	FARE MEDIA	2016	2015 *	2014	2013	2012	2011	2010	2009	2008
WHO	**ADULT**									
	TOKENS	102,073	110,945	111,157	112,360	117,962	124,748	120,366	114,686	94,210
	TICKETS	N/A	N/A	N/A	N/A	N/A	N/A	1,298	8,807	34,445
	TWO-FARE	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	PRESTO - SINGLE RIDE	27,397	13,323	9,862	8,194	4,399	1,139	0	0	N/A
	PRESTO - SRVM TOKEN RIDE	1,157	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	PRESTO - SRVM CASH RIDE	582	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	REGULAR MONTHLY PASS	194,820	204,509	214,932	213,982	205,086	194,928	203,101	208,172	203,313
	POST-SECONDARY PASS	51,861	48,396	42,855	38,426	35,019	32,091	9,200	N/A	N/A
	TWIN-GO PASS	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	WEEKLY PASS	7,547	8,843	9,361	9,557	10,185	9,893	9,237	8,738	7,517
	CASH	41,536	48,873	49,120	48,623	46,467	43,795	43,149	41,445	39,408
	SUB-TOTAL	**426,973**	**434,889**	******	**431,142**	**419,118**	******	**386,351**	**381,848**	******
	SENIOR/STUDENT									
	MONTHLY PASS	27,621	25,092	23,064	20,509	19,769	18,590	17,169	15,331	14,864
	WEEKLY PASS	959	672	515	540	624	702	814	874	780
	TICKETS	32,997	32,595	33,408	35,472	37,039	38,299	38,674	38,615	39,097
	TWO-FARE	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	PRESTO - SINGLE RIDE	1,421	438	12	N/A	N/A	N/A	N/A	N/A	N/A
	PRESTO - SRVM CASH RIDE	210	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	CASH	10,440	12,170	12,037	8,538	8,164	7,609	5,856	5,526	5,253
	SUB-TOTAL	**73,648**	**70,967**	**69,036**	**65,059**	**65,596**	**65,200**	**62,513**	**60,346**	**59,994**
	CHILDREN									
	FREE RIDES	21,875	10,939	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	TICKETS	0	1,066	7,097	7,563	7,929	8,304	8,287	8,562	8,782
	PRESTO - FREE CHILD RIDE	36	10	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	CASH	0	526	3,705	2,708	2,589	2,433	2,539	2,410	2,253
	SUB-TOTAL	**21,911**	**12,541**	**10,802**	**10,271**	**10,518**	**10,737**	**10,826**	**10,972**	**11,035**
	DAY/VIST./OTHER	9,130	8,561	10,033	11,428	11,929	10,642	10,605	10,880	9,961
	BLIND/WAR AMPS	1,088	1,086	1,119	1,109	1,086	1,060	1,073	1,074	1,092
	PREMIUM EXPRESS	474	490	451	401	372	344	322	313	310
	POSTAL CARRIERS	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	GTA PASS	4,855	5,471	6,087	5,784	5,388	5,642	5,667	5,800	5,415
	SYSTEM TOTAL	**537,476**	**524,005**	**534,435**	**525,164**	**514,697**	**500,219**	******	**471,235**	******
WHERE	**BUS**									
	BUS	252,899	238,943	245,292	239,968	234,582	223,269	219,855	218,545	215,997
	SUB-TOTAL	**252,899**	**238,943**	******	******	******	******	**219,855**	**218,545**	**215,997**
	RAIL									
	SUBWAY	221,622	228,129	219,849	217,250	216,101	213,280	199,131	199,321	196,004
	S.R.T.	2,951	3,352	4,254	4,661	4,667	4,766	4,232	4,300	4,639
	TROLLEY COACH	0	0	0	0	0	0	0	0	0
	STREETCAR	60,607	63,581	65,420	63,315	58,657	58,904	54,139	49,067	50,060
	SUB-TOTAL	**285,180**	**295,062**	******	******	******	******	******	******	******
	SYSTEM TOTAL	**538,079**	**534,005**	**534,415**	**529,194**	**514,997**	**500,216**	******	**471,232**	******
WHEN	WEEKDAY	424,117	423,808	423,269	416,297	406,913	395,578	379,810	374,908	374,765
	WEEKEND/HOLIDAY	113,962	110,197	111,546	108,897	107,094	104,641	97,547	96,325	91,935

The below table describes the details of each attribute in our data set.

Name	Description	Comments
2016 – 1985	This matrix shows the number of riders recorded under each column from 1985 to 2016. Each figure should have three zeros after each number.	Add three zeros to the end of each number. 112,360 is actually 112,360,000 tokens received from riders.
Fare Media	This lists the different forms of fares (payment) accepted by the TTC.	Tokens, tickets, PRESTO, monthly/weekly passes, cash
Who	This lists the different types of types of fares payable by different groups of riders	Adult – Tokens, Tickets, PRESTO, Regular Monthly Pass, etc. Children – Tickets, PRESTO, cash
Where	This identifies the types of vehicles being used to transport riders	Bus, subway, SRT, streetcar
When	This indicates by year the number of riders during a weekday or weekend/holiday	

2.3 Data Quality

The data quality is reasonable for the purpose. The data spans years 1985 through to 2016. It contains the ridership by passenger type, ticket type payment, and transportation (streetcar or subway). Data is current and up-to-date as of January 25, 2017.
Most useful and interesting was the classification of passenger types and the start of the Presto system.

3.0 Data Preparation

The following section discusses how we prepare and organize the data, cleaning out rows and columns that are not relevant and how we handled missing data. The following features were used in this project.

1. **sqlalchemy** package to save and read data from SQLite database
2. **numpy**
   round(), mean(), arrange(), array(), matrix()

3. **panda**
   concat(), read_sql_query(), read_excel(), read_csv(), autocorrelation_plot() pandas.plotting.table() draw data table in graph

4. **dataframe**
   set_index(), reset_index(), sort_index(), shift(), drop(), ffill(), fillna(0), astype(), to_sql(), min(), max(), sum(), mean(), groupby(), head(), tail(), transpose()

5. **str**
   format(), replace()

6. **pyplot**
   plot(), subplots(), acorr(), scatter() savefig()

7. **sklearn**
   LinearRegression(), fit(), score(), predict()

8. **statsmodels**
   ARIMA() model, fit(), forecast(), mean_squared_error()

9. **language**
   For loop, immediate if conditions (True if len(df)>0 else False)

3.1 Data Organization

Real life data is never perfect. There were rows, cells and columns that either had no data or contained information that were not relevant to our analysis. In addition, the TTC Ridership csv. file was formatted for presentation purposes therefore certain rows were left blank simply for ease of readability.

To effectively use our main data set, we had to determine how we needed to organize the data. This means renaming columns that do not make sense, adding in columns that were needed as part of our analysis and removing rows such as sub-totals that were no longer needed. We went through the cleanup process in first run then stored clean & processed data in our SQLite table. From the next run forward, data was retrieved from this table directly and there was no need to follow the data preparation process.

3.2 Data Cleanup

Now that we have determined how we want to organize the data, we proceeded with the following for our data clean-up.
The functions we used to clean up the data include:

- Dataframe rename functionality
  - Renaming columns to add clarity

- Dataframe drop functionality
    - Dropping rows that were not needed such as sub-totals
- Set row values to “who, what, where”
- Create new attributes 'Passenger' of passenger types, transportation, service days
- Used ffill() method to handle NAN values,
- Converting values to int by using .astype() method

Furthermore, we needed to address missing data in our data set. This is discussed below.

3.3 Handling Missing Data

Unfortunately, there were quite a few missing data within the datasets. Part of the reason why we have missing data is due to new programs being introduced or old programs being discontinued.

For this reason, it made our decision on selecting the method of handling missing data simple. We could not simply remove the rows as it would remove a large portion of information that is important.

In addition, with the removal of tickets and tokens and the introduction of Presto, keeping the data allows us to see how the trend move with regards to form of payment.

As such, we have determined that it was necessary to use the fillna() function to fill the NAN with 0’s.

The chart below highlights which fill method we selected for what data and why we selected that method:
For example, in this scenario we:

Fill Method	What	Why
Back fill [bfill]	Not used	
Forward fill [ffill]	The re-titled Ridership column included data which indicates ‘WHO, WHERE, WHEN’. Only the first row	We selected the forward fill method to fill this because the first row was representative for all rows that followed.

	for each type were indicated and the rest were left blank.	
Remove Data [fillna]	We had many rows that included NaN values due to Faremedia that were discontinued, not relevant or a new introduction to the TTC.	In these situations, instead of completely removing the row, we selected to use fillna because it still added value to see the method of payment where data existed

3.4 Data Transformation

After cleaning up the data, we focused on transforming the data for our analysis. We will now discuss the data transformation done in our analysis.

Since we had to drop rows we deemed as unnecessary, we used the Pandas reset_index function to reset the index. This was done for the TTC Ridership file, peak vs non-peak and Ridership Revenue.

Finally, for the Ridership Revenue file, the revenue was formatted as a string with a dollar value sign and comma's to break out the thousands (i.e. $53,000,000). To make it easy for our analysis, we transformed the data by doing the following:

- Remove the dollar sign and comma in the dollar amount
- Convert the revenue into millions so that we can remove 6 digits out of the number (1,000,000)
- Then we used the astype(int) function to convert the string into an integer
- Transpose the data to better illustrate the time series model

3.5 Data Storing and Verification

The processed data was stored in SQLite table and we ran the sql query on a trial basis to see if the data was transformed appropriately. As part of our verification – we used a few small sample datasets from the tables and matched it with the value produced from our raw dataset manually.

4.0 Data Analysis

The section provides our analysis of the data.

Before we begin our analysis, here is a snapshot of what the ridership make-up looks like presently (2016).

Who	Main pay method (2016)
Adults	Monthly pass
Children	Free rides
Senior/Students	Tickets followed closely by monthly pass

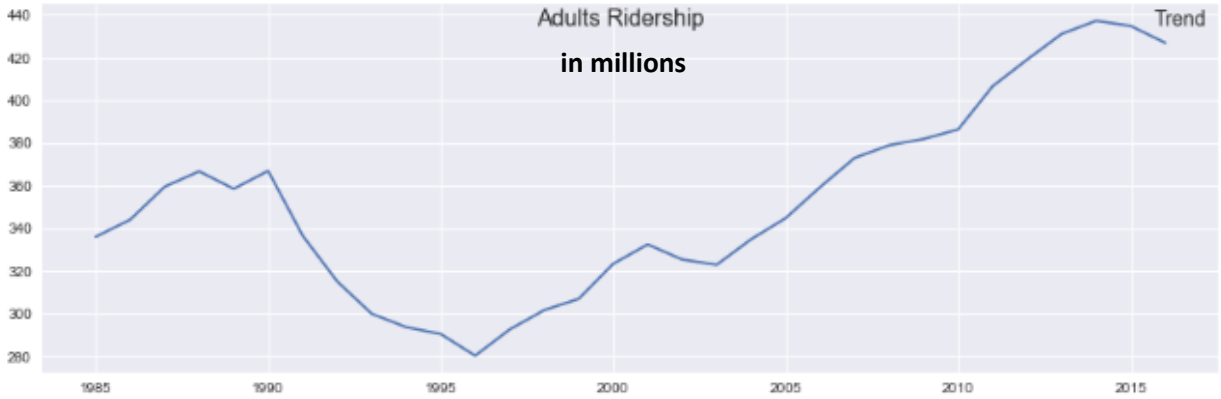
Type of transit	Percentage
Subway	77%
Street car	22%
SRT	1%

4.1 Trend Analysis

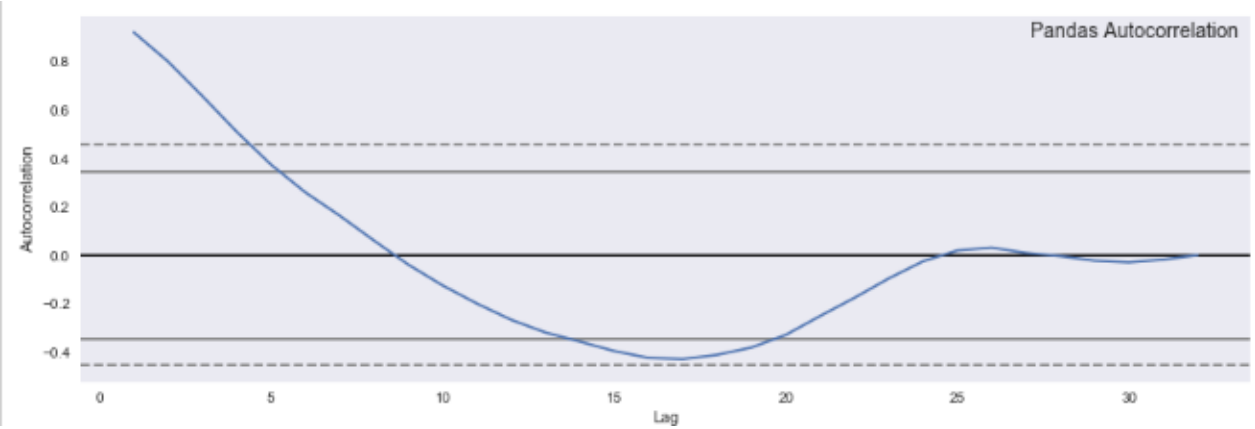
An analysis on the Passenger Type

Adults:

We can see that the trend for Adult ridership has been going up since 1996.

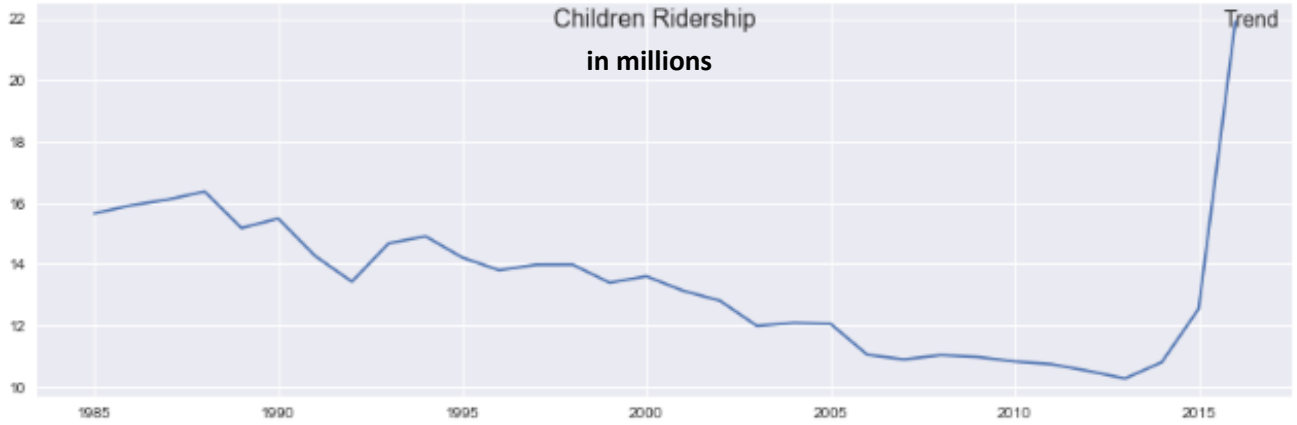


When looking at the chart, we find there is a statistically significant autocorrelation within a 5-year period.

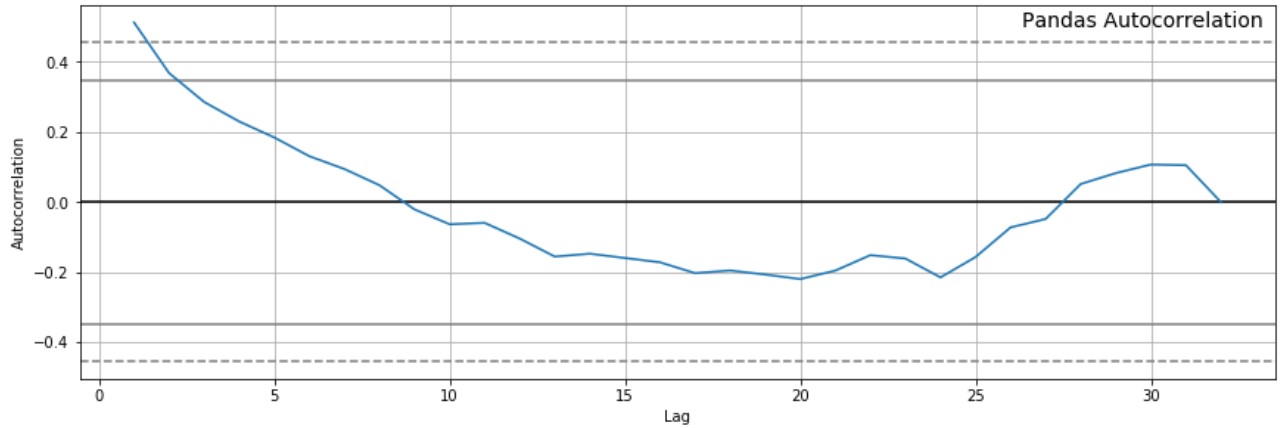


Children:

In 2015, a new rule was applied where Children under the age of 12 can ride the TTC for free. As a result, we can see a huge spike in 2015.

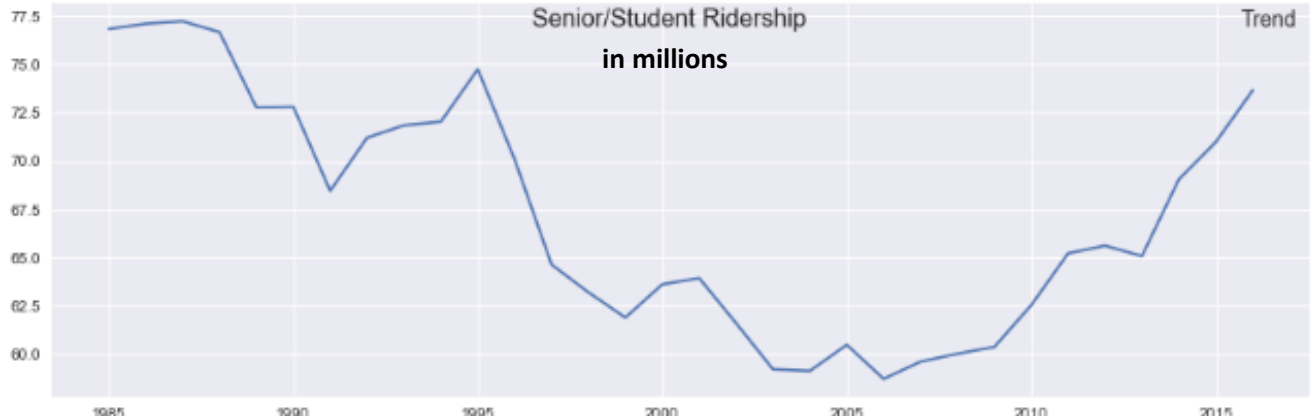


Due to the change in the rule, we don't see a high level of autocorrelation as displayed in the chart below.

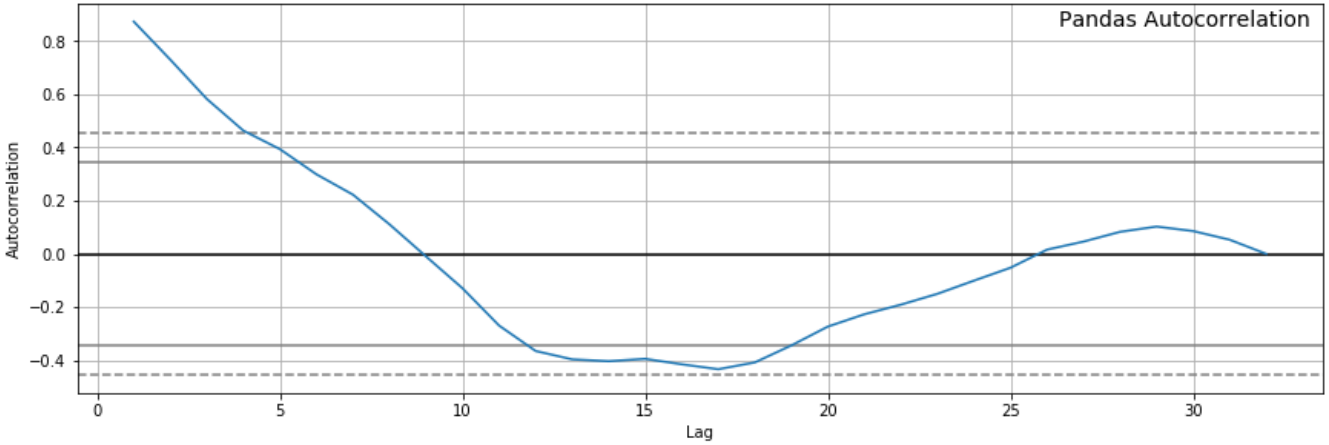


Seniors & Students:

There was a big dip from the early to late 2000's but it started picking up again by 2010.

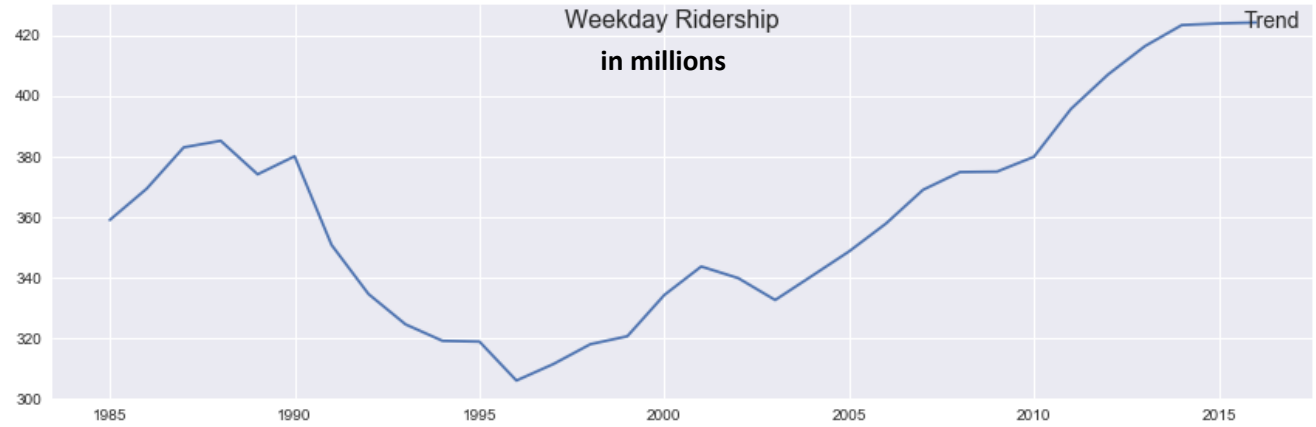


Like Adults, Students and Seniors have a high autocorrelation for about 5 years.

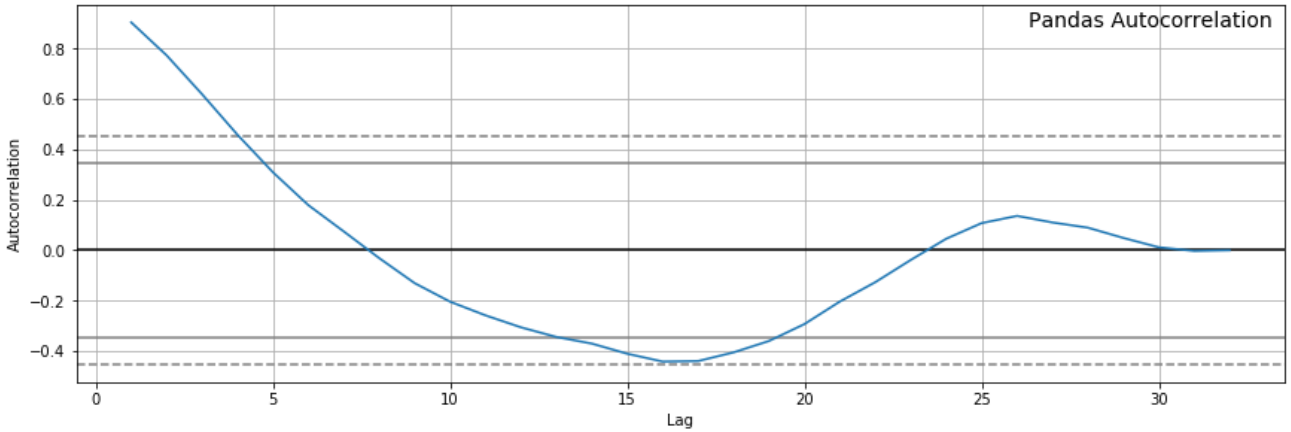


Weekday Ridership:

We see the trend of weekday ridership drop in 1990 possibly due to recession but began to pick up again in 2004.



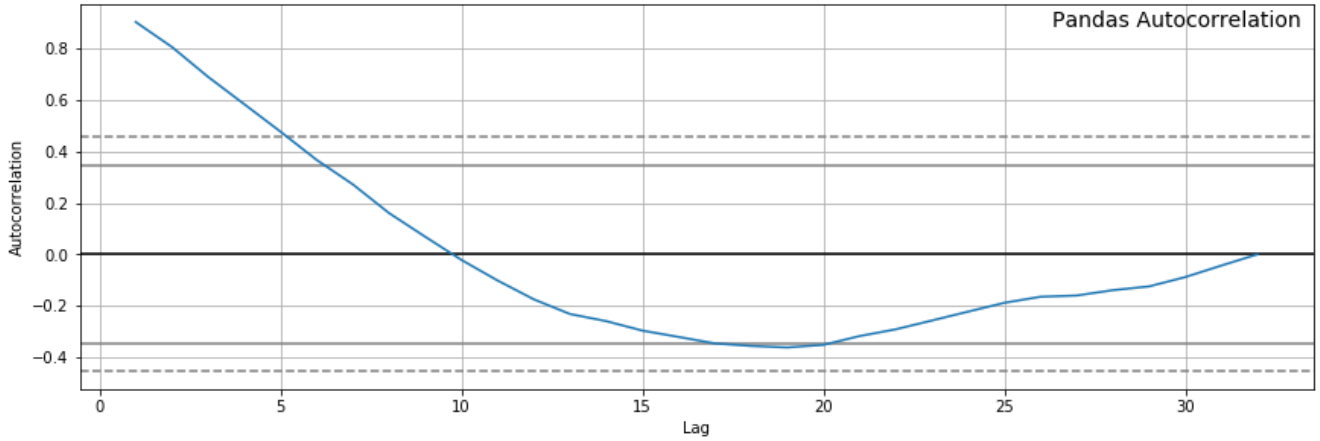
There is close to a 5-year autocorrelation in weekday ridership.



Weekend/Holiday Ridership:

The drop in 1994 in ridership was not as evident in the weekend vs the weekday numbers. However, the trend is similar where beginning in 2004, the trend picks up significantly.

The autocorrelation is higher in the weekend ridership at about 5 years vs the weekday numbers.

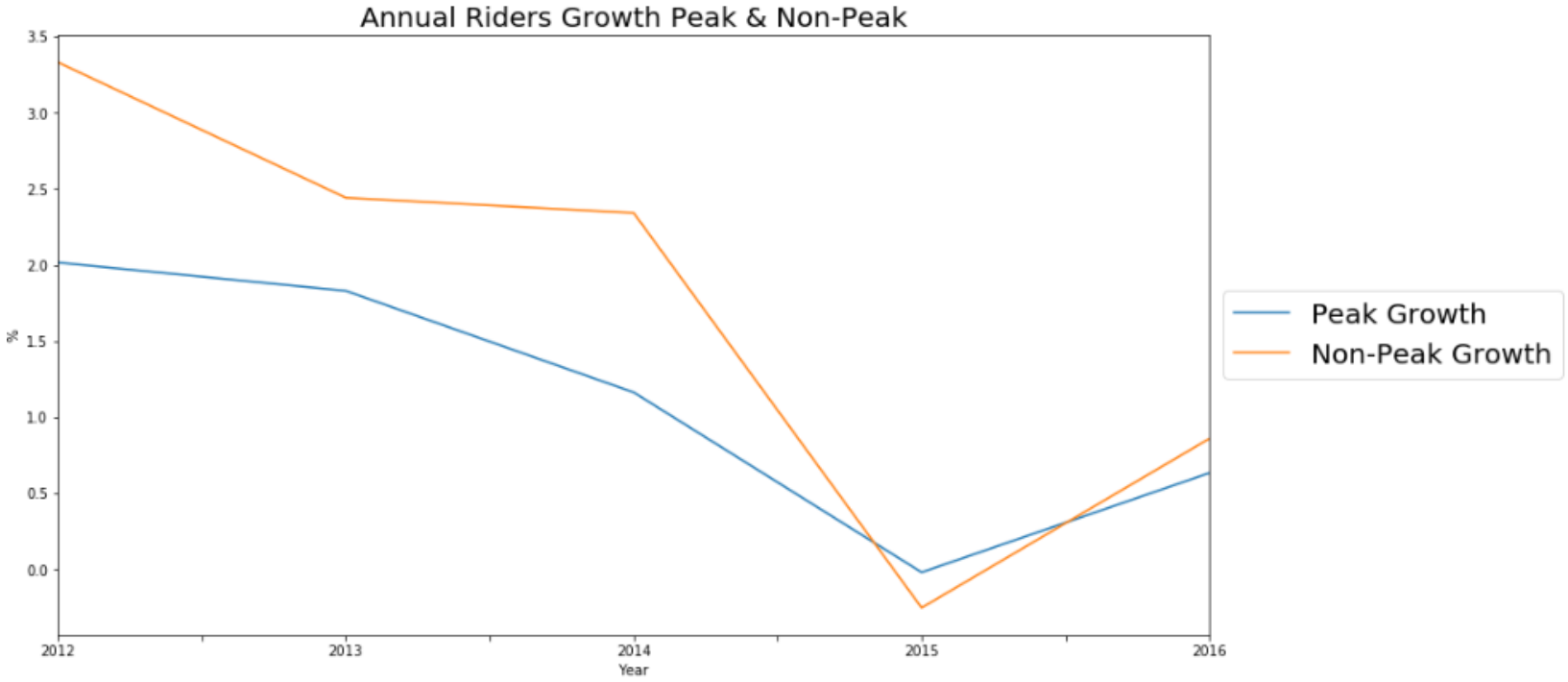


Monthly Passenger Rides

There is a strong seasonal trend in ridership where we see March, June, September and December is high. We suspect this is related to when kids are off school and parents are taking their kids out. In addition, September always show a peak. In this scenario, we believe it is related to back-to-school traffic. In addition, our research found that the last month in every financial quarter has 5 weeks instead of 4 which skews the data.

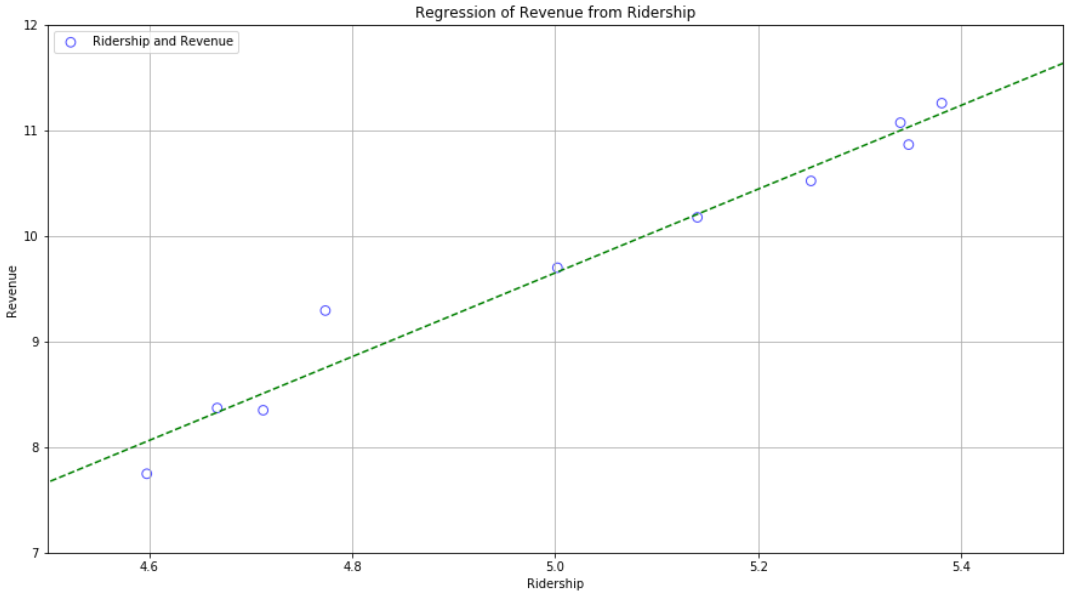
Peak vs. Non-Peak hours

The growth trend for Peak vs. Non-Peak are very similar. It was quite high in 2012 but both dropped dramatically in 2015 and started picking up again thereafter. The drop in 2015 was alarming and it was due to a slowing economy and employment as well as a fare increase. In addition, lower gas prices, the increase in popularity of Uber and allowing passengers to enter at the rear of streetcars who may not honour the system may play a role in the downward decline in 2015.



4.2 Regression Analysis

We explored our knowledge on linear regression model to depict the relationship of Revenue vs Ridership. It shows a linear relationship with a score of 0.96 (Revenue= -10.19 + 3.96 * Ridership).



OLS Regression Results

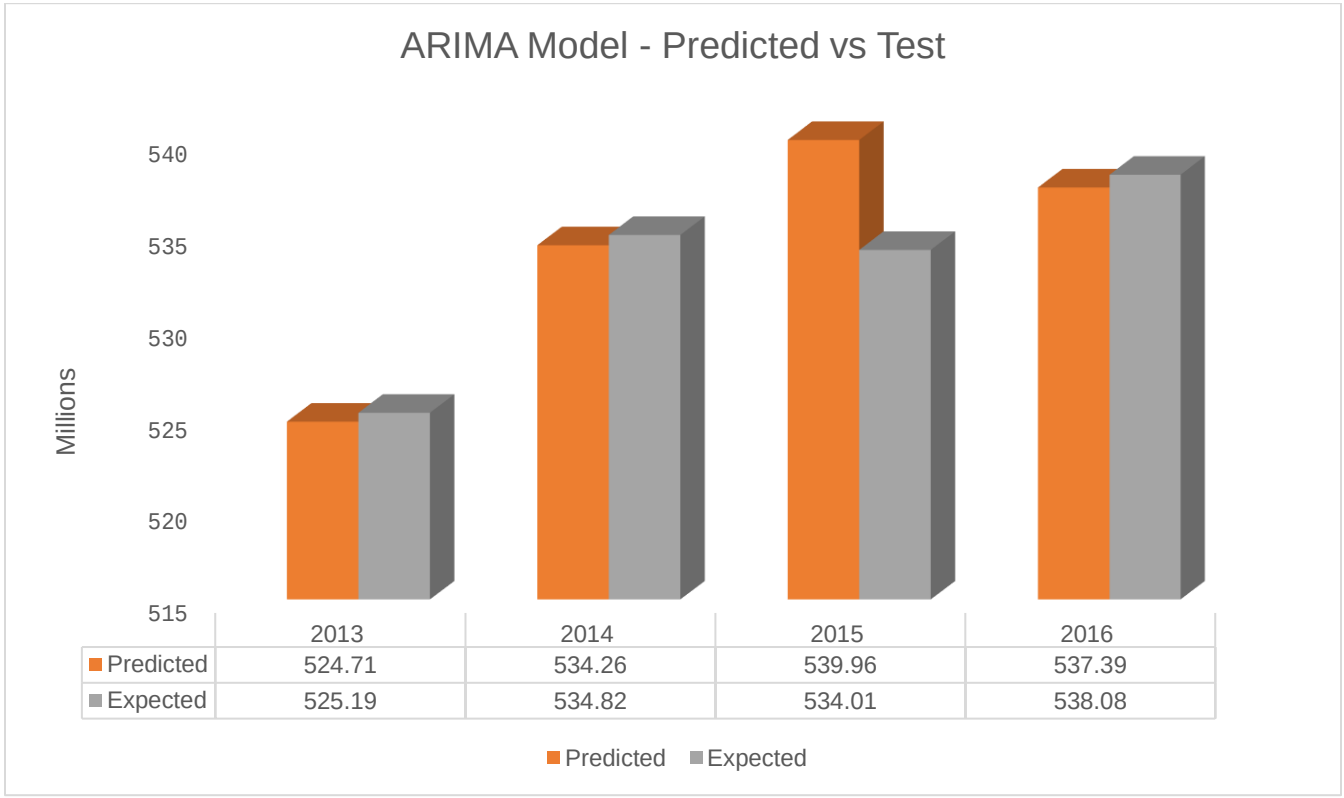
Dep. Variable:	Revenue	**R-squared:**	0.966
Model:	OLS	**Adj. R-squared:**	0.962
Method:	Least Squares	**F-statistic:**	229.5
Date:	Thu, 07 Dec 2017	**Prob (F-statistic):**	3.57e-07
Time:	01:29:39	**Log-Likelihood:**	1.0352
No. Observations:	10	**AIC:**	1.930
Df Residuals:	8	**BIC:**	2.535
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-10.1918	1.318	-7.735	0.000	-13.230	-7.153
Ridership	3.9684	0.262	15.150	0.000	3.364	4.572

Omnibus:	7.642	**Durbin-Watson:**	2.055
Prob(Omnibus):	0.022	**Jarque-Bera (JB):**	2.898
Skew:	1.158	**Prob(JB):**	0.235
Kurtosis:	4.260	**Cond. No.**	89.3

4.3 Data Modeling

We use ARIMA (Autoregressive Integrated Moving Average) Model for analyzing and forecasting our timeseries data. It explicitly caters to a suite of standard structures in time series data, and as such provide a simple yet powerful method for making skillful time series forecasts. We used 90% of data instead 66% of data as training data set and the rest of the data set was used to test the model. This model shows better result only about 9% of mean squared error.



This model helped us forecast the ridership to help the TTC plan their daily operational schedule. The TTC can use it to forecast their ridership in terms of passenger type, vehicle type and the traffic for weekday vs weekend/holiday.

5.0 Tools used

The following was used to prepare the data for our analysis:

Data perusal: Excel

Data Manipulation: Anaconda, Jupyter Notebook, Python's Pandas, Matplotlib, SKlearn and Numpy Libraries.

Group Collaboration: Atlassian's BitBucket (github) was used as our repository for sharing documentation and code.

6.0 Challenges

The following highlights the challenges we experienced when completing our project.

1) **Determining the dataset to use:** There are a lot of data available online, having to research to find a data set was one challenge. Each member found a dataset and the next challenge was to convince and persuade, and come to an agreement on which was the best to use
2) **Multiple files**: We felt that one data set was insufficient for our analysis. We had to use 4 different files to complete a more holistic picture of the TTC ridership analysis. The files we used were:
    a. TTC Ridership
    b. Revenue
    c. Peak hours
    d. Non-peak hours
3) **Determine how to format and display the data:** Figuring out how we should display the data and what time frame makes sense. Other factors include determining what columns are relevant, what is not and what new columns we needed to add. In addition, the data clean up fill method were considerations as well.
4) **Coding:** Another challenge was the coding of the notebook
5) **Incorporating our analysis with research**. The trend shows us numbers and in some instances, we needed to do additional research to figure out reasons why.
6) **What to apply from class:** Determining what which component and section we have learned from class to apply to our analysis

7) **Graphs**: Determining which graph to portray our data that will allow the user to easily comprehend the information and it is displayed in a useful layout.
8) **Conclusions**: Drawing conclusions on our data, figuring out how to interpret the data and make sense of it. Determining why the following months have higher revenue; March, June, September, December. Determine why September is always the peak?

7.0 Conclusions

Based on our research and analysis within this report, we see that the trend for passenger ridership for adults, children, senior/students have been going up. Interestingly, each group have a different dominant method of payment. We see there is a consistent high level of autocorrelation for approximately 5 years for each passenger type with the exception for children where free child fares have modified the trend significantly. We see an evident seasonal trend with ridership where there is a peak in March, June, September, and December.

To answer the question 'what should be the TTC's focus in development for 2018, we believe it needs to focus on Presto.

For 2016, Presto payment is still lagging significantly behind other forms of payment. The adoption of Presto should be a priority for the TTC especially with the group adults and students who use the TTC as a regular form of transportation. Presto allows a seamless form of payment for individuals who are traveling from city to city without having to switch fares. In addition, it is more cost effective than monthly passes which require the TTC to print new passes each month. The Presto card is issued to the passenger at $6 per card. With this cost, this encourages passenger to keep their card for a long time.

To increase Presto usage, the TTC need to continue to implement Presto charge stations at more locations. This will allow people to use the Presto card anywhere.

Presto will change the way people travel in the GTA and the TTC need to put an emphasis on developing its network in 2018.

8.0 References

Our data was retrieved from the TTC website:

https://www1.toronto.ca/wps/portal/contentonly?vgnextoid=a3b7c87477438310VgnVCM1000003dd60f89RCRD

https://www.ttc.ca/Fares_and_passes/PRESTO/PRESTO_Subway.jsp

https://www.prestocard.ca/en/shopping/order-a-card

https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

https://www.theglobeandmail.com/news/politics/drop-in-transit-ridership-has-officials-across-canadastumped/article30178600/

10.0 Appendices

1. Jupiter Notebook
2. Group Presentation
3. Datasets