CSP554 - BIG DATA PROJECT REPORT

CHICAGO PUBLIC TRANSPORTATION ANALYSIS

Suraj Nammi

Kruthi Kanukuntla

ABSTRACT:

The city of Chicago has the best public transportation and best public transportation service provided by The Chicago Transport Authority (CTA). CTA made public transportation more sophisticated and more user friendly for the riders in Chicago. All the buses and rails are tracked using application and information about the services are given live updates to the users in the mobile application. This project is mainly deals with data analysis and visualization. We use the data taken from CTA riders and perform data operations and make an analysis and answer few questions and draw statistical graphs showing how the data is changing timely. The purpose of this project is to show how CTA ridership evolved in the past two decades and draw conclusions on how to develop or make the CTA transportation more efficient for the future. This data analysis is done on the data which is verified by CTA officials and most part of the data is taken from CTA users. We also understand how Big Data Analytics is important for the organizations and how it plays an important role in decision making skills.

I. INTRODUCTION:

We collected the data from 2001 to 2022 of all the public transportation used by commuters in Chicago. We analyze and observe the trends in the data by representing them in graphs and charts and also try to create an analysis where we can see the CTA buses and rails data from different time periods and find out observations on different days and different seasons. To complete this project, we are working on two different datasets and work on it. This project has the potential to use different data visualization tools and perform different approaches to get the best results and outputs. The data is represented in forms of graphs and a detailed explanation of every graphical representation will be given. The project also yields solutions to different problems that are potential in near future of CTA. The project is performed using PySpark, we used PySpark for faster and more interactive results. PySpark supports SQL and Python and allows users to create data frames over the dataset. Creating dataframes will make it more optimized to perform analytics.

PROJECT TOOLS:

- Apache Spark
- Jupyter
- Python

II. COMPARISON OF TOOLS:

Scala

- The Scala and Python APIs are both excellent for the majority of tasks,and Spark is a fantastic framework.
- Powerful programming language Scala provides developer-friendly features that Python does not. Many machine learning methods, including regression, classification, clustering, and decision trees, to mention a few, are already well supported by Spark.
- There are a few factors to take into account, though for using Python over Scala for Spark:
    - Many data science libraries are available in Python that can be connected with PySpark. Pandas, a Python library for doing numerical and statistical analysis, is built on top of NumPy arrays.
    - Python includes libraries like Matplotlib, TensorFlow, SciKit Learn, SciPy, and Statmodels for the purposes of data science approaches. Breeze and ScalaNLP are available in Scala for less sophisticated numerical algorithms. However, it is discovered that the Python libraries offer a better scope when compared to the same libraries offered by Scala language.
    - Data scientists can concentrate more on the data science aspect of their code rather than the syntax of the code to satisfy their objectives because Python also has a less difficult learning curve and is simpler to create.

Pyspark

- To assist us control Big Data, PySpark API combines the practicality of Python with the strength of Apache Spark. Spark is a complete project that has produced an entire ecosystem. It fits really well with the project we were trying to create.
- PySpark is superior to standard Python machine learning applications in that it offers comprehensive and high-quality methods for processing zetabytes and petabytes of data on parallel clusters much more quickly. A machine learning pipeline and statistical analysis techniques are also included in PySpark.
- Python is an easy language to learn and implement. It provides a simple yet comprehensive API. With Python, the code is far more readable, and maintenance and familiarity is far better. In contrast to Java or Scala, it offers many options for data visualization which we can take advantage of, by using PySpark.
- In addition to its simplicity and its ability to handle errors, its ease of use makes Pyspark the best technology for CTA dataset analysis.

Hive

- Apache Hive and Apache Spark are the two popular Big Data tools available for complex data processing.
- Hive is built on top of Hadoop and provides the measures to read, write, and manage the data. HQL or HiveQL is the query language in use with Apache Hive to perform querying and analytics activities.
- Hive, for instance, does not support sub-queries and unstructured data. It is also not a suitable choice for real-time online transaction processing applications. Data update and deletion operations are also not possible with Hive.
- Spark SQL supports real-time online transaction processing along with row-level updates. These features are not present in Apache Hive.

III. DATA OPERATIONS AND PROCESS FOR ANALYSIS:

- ***Data processing:*** We processed the data into required fields using python. We created a spark session and read the dataset (CSV's). After reading the data, we are using PySpark for data profiling the data to be used for data analysis. The data is mostly preprocessed by CTA for data integrity.
- ***Data Integration:*** We integrated two different datasets to show the analysis of CTA ridership. We read the data from the csv file took from the CTA official website in Jupyter Notebook using python. The data we are integrated has data from 2001 till 2022. We are extracting data year wise, month wise and even day wise for further analysis. We are using two different datasets
- ***Data Profiling:*** We performed data operations to change the formats of Date to read them in Spark. We replaced string format columns to Integer and float formats to perform mathematical and logical operations. We renamed columns in the dataset to perform data operations and perform analysis. We replaced null values with min averages and mean of data to get the most accurate results.
- ***Data Visualization:*** We created Interactive Charts and Graph to represent the historical data and develop an analysis for CTA routes that are the busiest and have large commuters. We used plotly to plot the graphs. We created dataframes using PySpark and used those to plot graphs for the analysis.
- ***Data Analytics:*** We performed different operations on the fields to get the required analysis. We studied the data and transformed data into graphs. We studied the graphs and made analysis of all the data.

IV. DATASET EXPLANATION:

The Dataset we used in the project is directly taken from Chicago City Dataset Organization. We used datasets focusing on CTA Boardings for the analysis of commuters using CTA as their mode of transportation. Both the datasets together have 2,86,409 records. The datasets have different types of records like integer, float, strings, dates, etc., we used every column and performed data operations to change the data types and perform operations. The CTA ridership dataset attributes have the information on a daily basis. CTA Bus Routing dataset attributes have the information on a monthly basis. We have records from 2001-2022(until July) in CTA ridership dataset and CTA Bus Routing Dataset. The datasets are also processed by CTA Authority for the fact checks and then made available for developers. This makes our Data analysis more factual and accurate to reality.

CTA Ridership Dataset has following attributes:

- ***Service_date:*** Records of dates on which CTA is on service.
- ***day_type:*** Records of type of the day, W = Weekday, A = Saturday, U = Sunday/Holiday.
- ***Bus_Boardings:*** Records of commuters travelling on CTA Bus on the given service_Date and Day_Type.
- ***Rail_Boardings:*** Records of commuters travelling on CTA Railway(Metro) on the given service_Date and Day_Type.
- ***Total_rides:*** Records of total commuter travelling on CTA Bus/CTA Rail are taken by Summing the Bus_Boardings and Rail_Boardings.

CTA Route Dataset has following attributes:

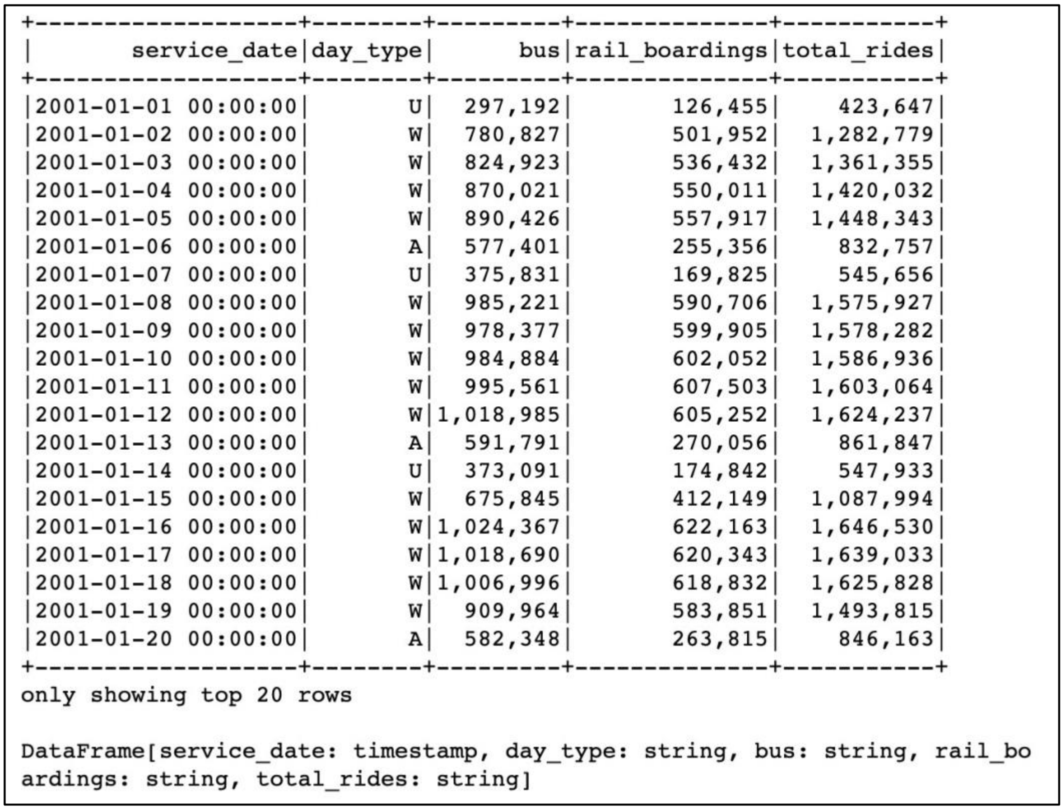
- ***Route:*** Records of the CTA Bus routes
- ***Route Name:*** Records of The Route Names of the CTA Buses
- ***Month_Begining:*** Records of Dates of services of CTA Buses
- ***Avg_weekday_Rides:*** Records of averages of weekday rides
- ***Avg_Saturday_Rides:*** Records of averages of Saturday rides
- ***Avg_Sunday_Rides:*** Records of averages of Sunday rides.
- ***Avg_Total_Rides:*** Records of sums of averages of weekday rides, Saturday rides and Sunday rides.

These attributes are processed in the PySpark module, and we performed CURD operations on the datasets. Using Pyspark we also performed operations like groupby, filter, sort for analysis of the data.

V. Observations and Analysis:

CTA Ridership Dataset:

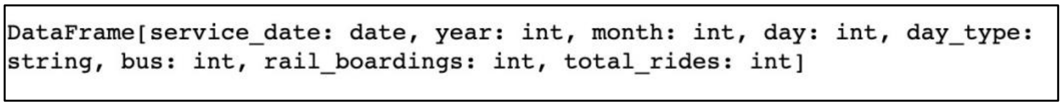
We can see the top 20 rows of the dataset with attribute headers and its schema just after loading it through the file.



```
+-------------------+--------+---------+--------------+-----------+
|       service_date|day_type|      bus|rail_boardings|total_rides|
+-------------------+--------+---------+--------------+-----------+
|2001-01-01 00:00:00|       U|  297,192|       126,455|    423,647|
|2001-01-02 00:00:00|       W|  780,827|       501,952|  1,282,779|
|2001-01-03 00:00:00|       W|  824,923|       536,432|  1,361,355|
|2001-01-04 00:00:00|       W|  870,021|       550,011|  1,420,032|
|2001-01-05 00:00:00|       W|  890,426|       557,917|  1,448,343|
|2001-01-06 00:00:00|       A|  577,401|       255,356|    832,757|
|2001-01-07 00:00:00|       U|  375,831|       169,825|    545,656|
|2001-01-08 00:00:00|       W|  985,221|       590,706|  1,575,927|
|2001-01-09 00:00:00|       W|  978,377|       599,905|  1,578,282|
|2001-01-10 00:00:00|       W|  984,884|       602,052|  1,586,936|
|2001-01-11 00:00:00|       W|  995,561|       607,503|  1,603,064|
|2001-01-12 00:00:00|       W|1,018,985|       605,252|  1,624,237|
|2001-01-13 00:00:00|       A|  591,791|       270,056|    861,847|
|2001-01-14 00:00:00|       U|  373,091|       174,842|    547,933|
|2001-01-15 00:00:00|       W|  675,845|       412,149|  1,087,994|
|2001-01-16 00:00:00|       W|1,024,367|       622,163|  1,646,530|
|2001-01-17 00:00:00|       W|1,018,690|       620,343|  1,639,033|
|2001-01-18 00:00:00|       W|1,006,996|       618,832|  1,625,828|
|2001-01-19 00:00:00|       W|  909,964|       583,851|  1,493,815|
|2001-01-20 00:00:00|       A|  582,348|       263,815|    846,163|
+-------------------+--------+---------+--------------+-----------+
only showing top 20 rows

DataFrame[service_date: timestamp, day_type: string, bus: string, rail_bo
ardings: string, total_rides: string]
```

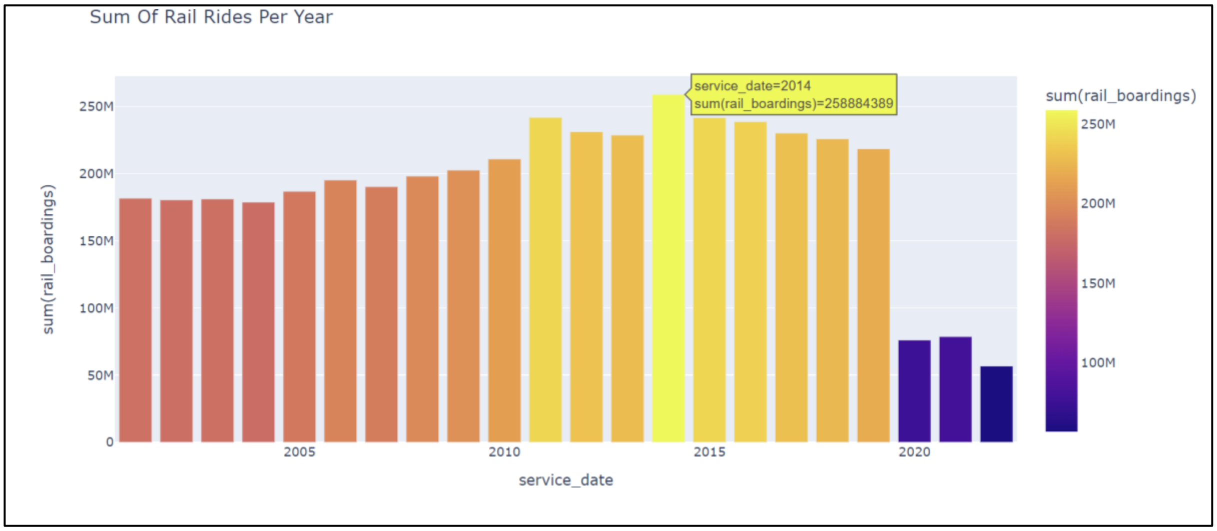
Schema of the dataset after we did couple of modifications to it.



```
DataFrame[service_date: date, year: int, month: int, day: int, day_type:
string, bus: int, rail_boardings: int, total_rides: int]
```

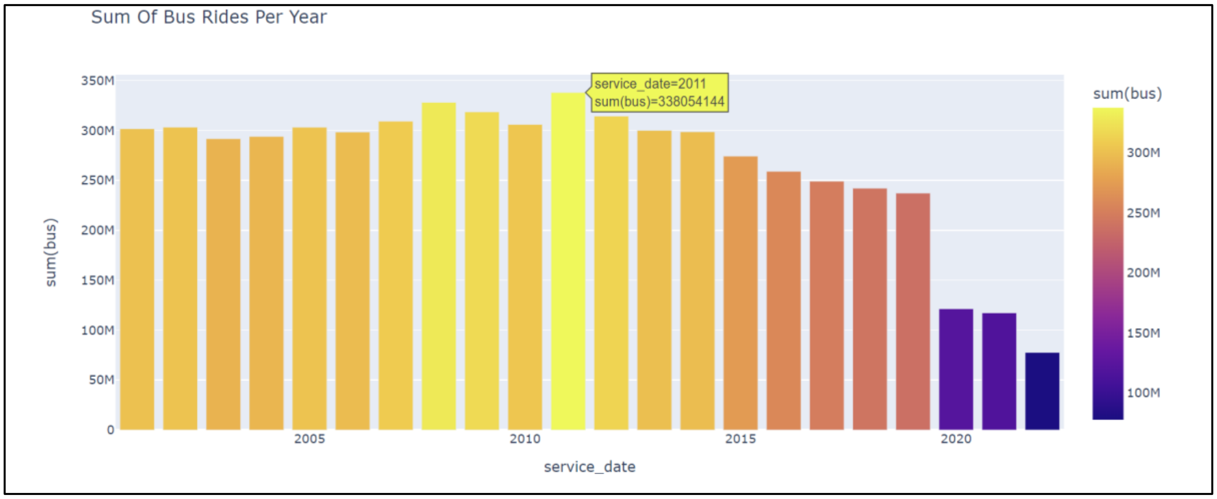
Visualization of the CTA Ridership Dataset in a Spark Session.

1. Sum of Rail Rides Per Year



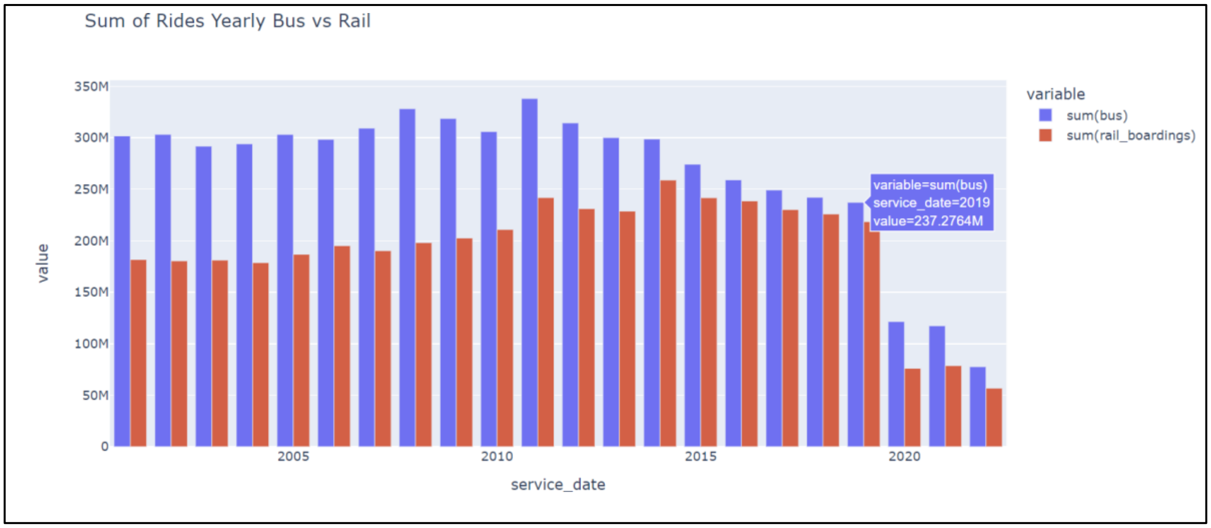
The data shows the analysis of the Rail Rides for every year from (2001-2022). We summed total number of rides for each year compared it to see how the ridership is changing yearly. We can see how the graph is peaking in the year 2014 to almost 250M. We can see a negative trend during the Covid years 2020 and 2021. We can see how the ridership took a heavy slump.

2. Sum of CTA Bus Rides Per Year



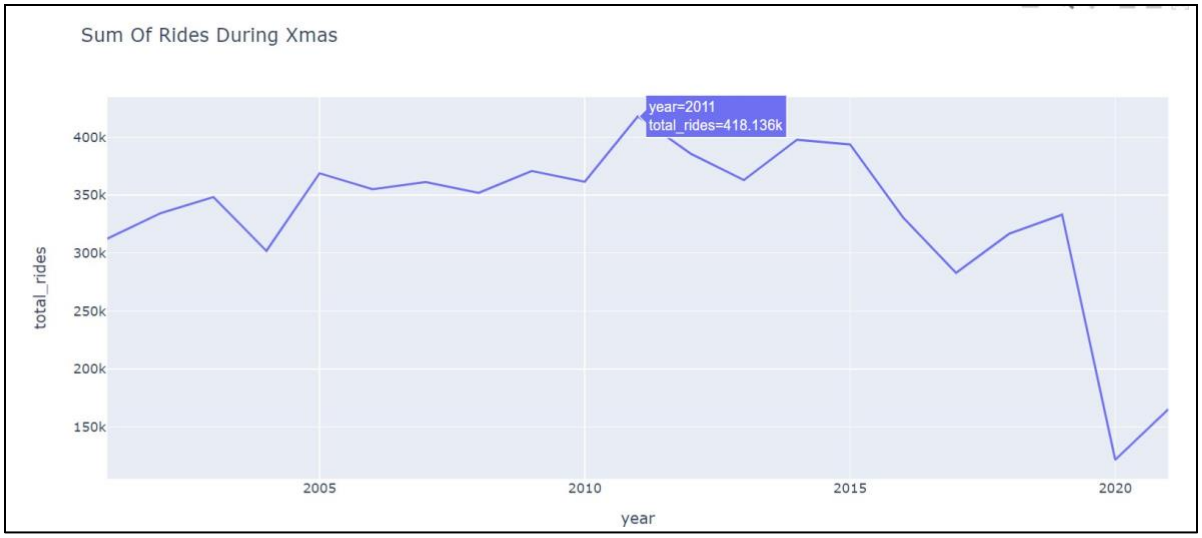
The data shows the analysis of the Bus Rides for every year from (2001-2022). We analyzed the bus rides here like the graph above. We can see how the graph is peaking in the year 2011 to almost 338M. We can see how the ridership took a heavy setback during the Covid years 2020 and 2021.

3. CTA Bus Vs CTA Rail Ridership



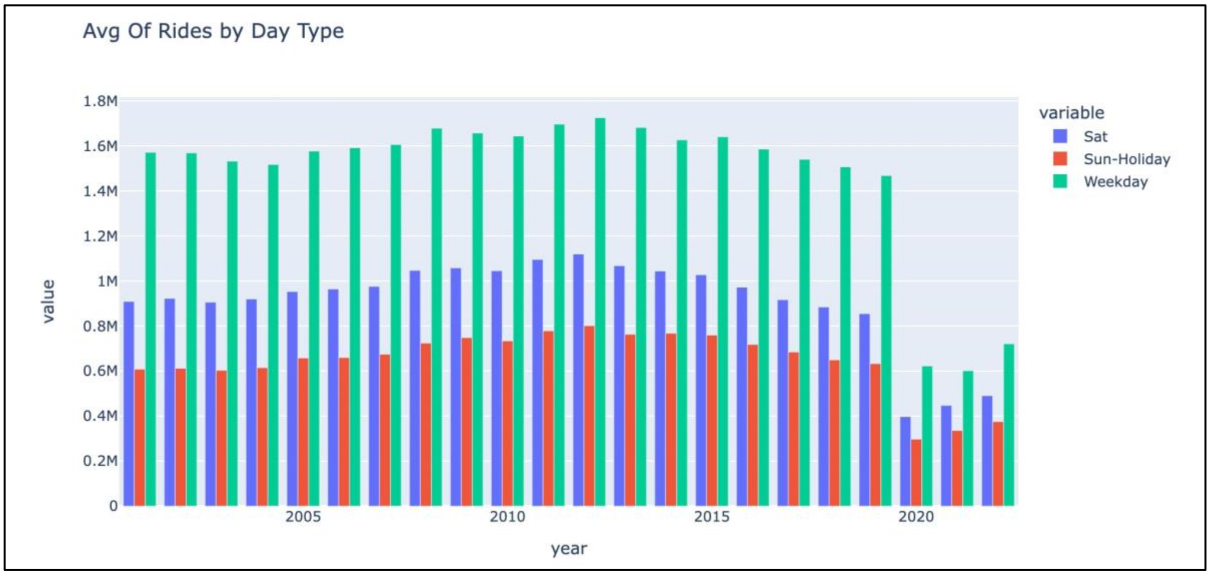
The data shows the analysis of the Rail Rides and Bus Rides for every year from (2001-2022). We summed total number rides and compared it yearly. We can see how the commuters using both CTA Bus and CTA Rails equally in their day-to-day. We can see how the ridership took a heavy debt during the Covid years 2020 and 2021.

4. CTA Bus and CTA Rail Rides During the Christmas EVE



The data shows the analysis of the Rail Rides and Bus Rides for every year from (2001-2022) on 25th December. We summed total number rides on Christmas day every year and compared it yearly to see how the ridership is changing year by year. We can see how the commuters using both CTA Bus and CTA Rails peaked in the year 2011 where we can see almost 418K people used the CTA Bus and Rail ridership's traveled on CTA that year. We can see how the ridership is at its all-time low compared to the past 2 decades.

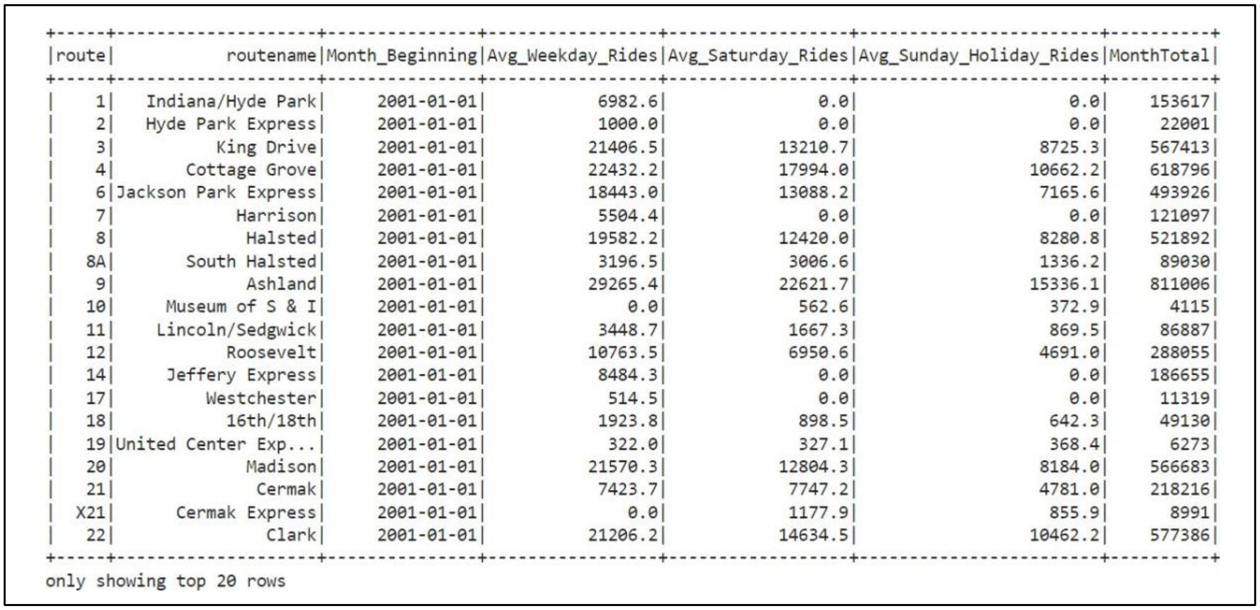
5. Average Ridership on Given Day Type



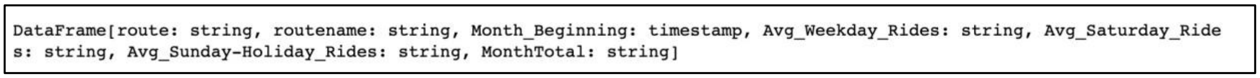
We can clearly see that Green Bars are having higher values compared to blue and red bars. Hence, we can easily draw a conclusion that the average ridership is more on Weekdays compared to Saturdays and Holidays. The CTA Transportation is busy on week days compared to weekends.

CTA Route Dataset:

We can see the top 20 rows of the dataset with attribute headers and its schema just after loading it through the file.

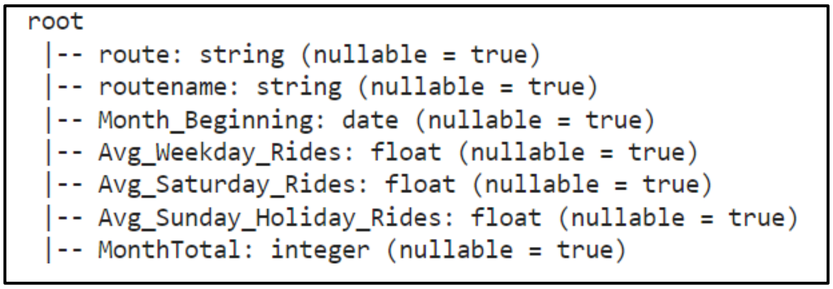


```
+-----+--------------------+---------------+-----------------+------------------+------------------------+----------+
|route|           routename|Month_Beginning|Avg_Weekday_Rides|Avg_Saturday_Rides|Avg_Sunday_Holiday_Rides|MonthTotal|
+-----+--------------------+---------------+-----------------+------------------+------------------------+----------+
|    1|   Indiana/Hyde Park|     2001-01-01|           6982.6|               0.0|                     0.0|    153617|
|    2|   Hyde Park Express|     2001-01-01|           1000.0|               0.0|                     0.0|     22001|
|    3|          King Drive|     2001-01-01|          21406.5|           13210.7|                  8725.3|    567413|
|    4|       Cottage Grove|     2001-01-01|          22432.2|           17994.0|                 10662.2|    618796|
|    6|Jackson Park Express|     2001-01-01|          18443.0|           13088.2|                  7165.6|    493926|
|    7|            Harrison|     2001-01-01|           5504.4|               0.0|                     0.0|    121097|
|    8|             Halsted|     2001-01-01|          19582.2|           12420.0|                  8280.8|    521892|
|   8A|       South Halsted|     2001-01-01|           3196.5|            3006.6|                  1336.2|     89030|
|    9|             Ashland|     2001-01-01|          29265.4|           22621.7|                 15336.1|    811006|
|   10|     Museum of S & I|     2001-01-01|              0.0|             562.6|                   372.9|      4115|
|   11|    Lincoln/Sedgwick|     2001-01-01|           3448.7|            1667.3|                   869.5|     86887|
|   12|           Roosevelt|     2001-01-01|          10763.5|            6950.6|                  4691.0|    288055|
|   14|     Jeffery Express|     2001-01-01|           8484.3|               0.0|                     0.0|    186655|
|   17|         Westchester|     2001-01-01|            514.5|               0.0|                     0.0|     11319|
|   18|           16th/18th|     2001-01-01|           1923.8|             898.5|                   642.3|     49130|
|   19|United Center Exp...|     2001-01-01|            322.0|             327.1|                   368.4|      6273|
|   20|             Madison|     2001-01-01|          21570.3|           12804.3|                  8184.0|    566683|
|   21|              Cermak|     2001-01-01|           7423.7|            7747.2|                  4781.0|    218216|
|  X21|      Cermak Express|     2001-01-01|              0.0|            1177.9|                   855.9|      8991|
|   22|               Clark|     2001-01-01|          21206.2|           14634.5|                 10462.2|    577386|
+-----+--------------------+---------------+-----------------+------------------+------------------------+----------+
only showing top 20 rows
```



```
DataFrame[route: string, routename: string, Month_Beginning: timestamp, Avg_Weekday_Rides: string, Avg_Saturday_Ride
s: string, Avg_Sunday-Holiday_Rides: string, MonthTotal: string]
```

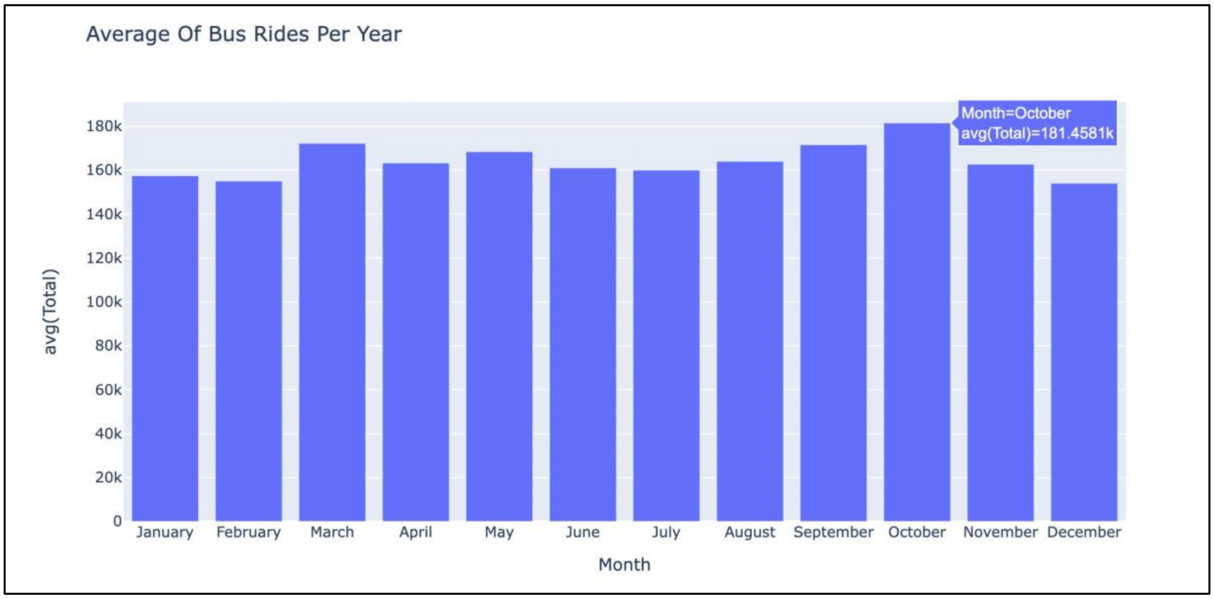
Schema of the dataset after we did couple of modifications to it.



```
root
 |-- route: string (nullable = true)
 |-- routename: string (nullable = true)
 |-- Month_Beginning: date (nullable = true)
 |-- Avg_Weekday_Rides: float (nullable = true)
 |-- Avg_Saturday_Rides: float (nullable = true)
 |-- Avg_Sunday_Holiday_Rides: float (nullable = true)
 |-- MonthTotal: integer (nullable = true)
```

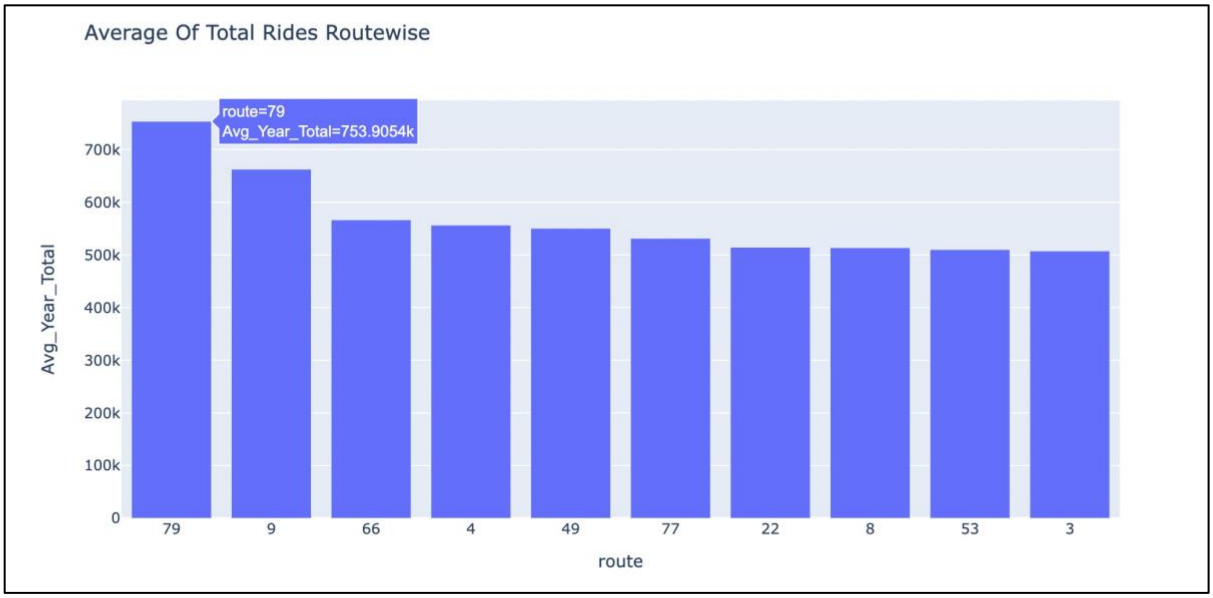
VISUALIZATION OF THE CTA ROUTE DATASET IN A SPARK SESSION:

1. Average Bus Rides Compared Yearly



The data shows the analysis of the Bus Rides for average ridership of every month from (2001-2022). The data is directly taken from Chicago City Datasets. We aggregated total number rides every day and compared it monthly to see how the ridership is changing month by month. We can see how the graph is peaking in the month October with an average ridership of 181.45K riders in the CTA Bus ridership's traveled on CTA. We can see how consistent the graph is and understand that an average of 150K riders is using CTA Bus every month.

2. Top 10 Busiest CTA Routes:



In the graph we can see all the top 10 busiest or most used CTA bus Routes over the past two decades. We can clearly see that Route 79 is the busiest CTA Bus route with average ridership of 753K riders. The graph also shows how CTA can improve their services in these routes better serve the commuters better.

VI. FINAL CONCLUSIONS ON DATA ANALYSIS:

We can clearly state that Covid-19 has affected the most on CTA Ridership. Covid years 2020-2021 lockdowns have slowed down the commuter from using CTA. This slump can also affect the CTA economy. The graphs also show how the passengers are worried about travelling in public transportation after the Covid-19. We can also see the busiest routes in Chicago with the help of CTA routes datasets and this analysis can be used by CTA to understand where they need to focus and give better service to the commuters on the busiest routes by increasing the frequency of buses on that respective route.

VII. FUTURE SCOPE

There is a lot of scope on the data analysis which we couldn't show in our project. Few analyses we found out that can be helpful in the near future are.

- Analysis of Traffic on the busiest CTA bus routes.
- Analysis of Economy Effects of Covid – 19 on CTA.
- Analysis of Rush hours and Busiest hours of CTA.
- Analysis for cost reduction by changing bus routes on basis of ridership.
- Analysis on how to increase the frequency of buses on workdays compared to weekends
- Analysis on how to develop or increase CTA riders in future. By forecasting the demand of buses and bus routes.

VIII. REFERENCES:

1] CTA Ridership Data: https://data.cityofchicago.org/Transportation/CTA-Ridership-Daily-Boarding-Totals/6iiy-9s97

2] CTA Ridership Buse Route Data: https://data.cityofchicago.org/Transportation/CTA-Ridership-Bus-Routes-Monthly-Day-Type-Averages/bynn-gwxy

3] Apache Spark: https://www.projectpro.io/article/scala-vs-python-for-apache-spark/213

4] Hive: https://www.projectpro.io/article/spark-vs-hive/480