



Data sets were downloaded from the linked website and easily merged using a dos command prompt due to the gigantic nature of the data. 2015 through 2018 was chosen because 2019 data was incomplete. 2013 and 2014 may have bad data and having not completely refined their process yet may not have a full data set. For sake of size and time and interest in comparing data over years, a very large data set of over 53 million records are analyzed below.

```
Administrator: Command Prompt
Microsoft Windows [Version 10.0.18362.418]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\WINDOWS\system32>cd\tableau

C:\Tableau>copy *.csv 2015.csv
201501-citibike-tripdata.csv
201502-citibike-tripdata.csv
201503-citibike-tripdata.csv
201504-citibike-tripdata.csv
201505-citibike-tripdata.csv
201506-citibike-tripdata.csv
201507-citibike-tripdata.csv
201508-citibike-tripdata.csv
201509-citibike-tripdata.csv
201510-citibike-tripdata.csv
201511-citibike-tripdata.csv
201512-citibike-tripdata.csv
        1 file(s) copied.

C:\Tableau>
```

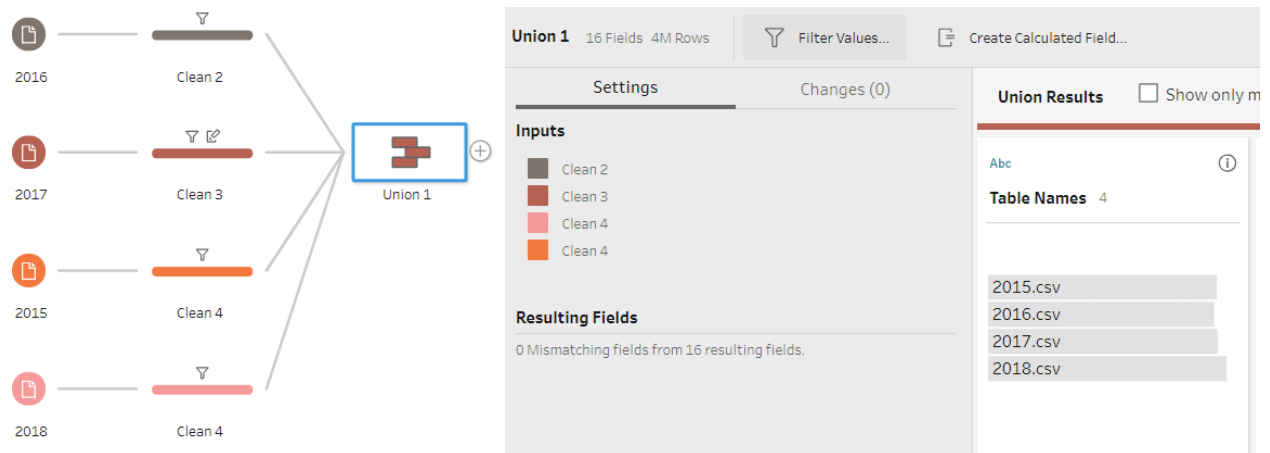
To complement the learning done in class, I completed the **LinkedIn Learning: Cleaning, Transforming and Prepping Your Data With Tableau Prep** course online in order to be skilled with using Tableau Prep Builder 2019.3 to clean such a huge dataset. I also followed along to **LinkedIn Learning: Creating Interactive Tableau Dashboards**

Key Filtering Steps:

Kept trip duration <25000 seconds.

Birth year > 1930 and removed null values

cleaned 2017 column names that needed to be renamed to align with 2015, 2016 & 2018 data so that the resulting fields would have no mismatching fields.



Removed latitude & longitude extreme outliers & labeled gender as Male, Female and Unknown.

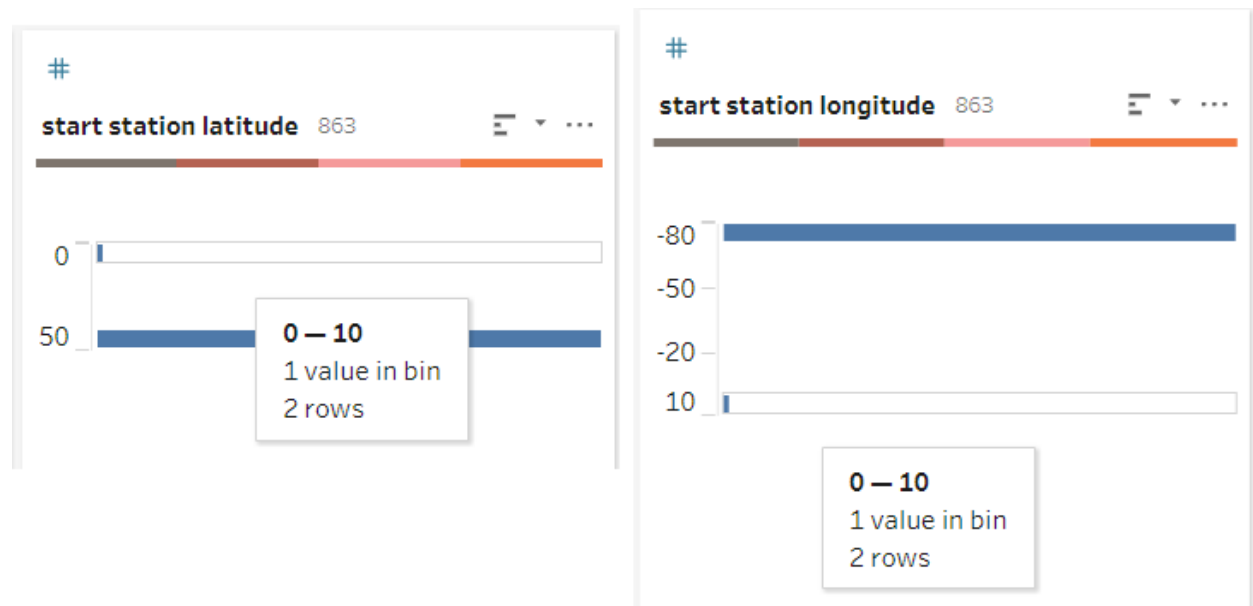
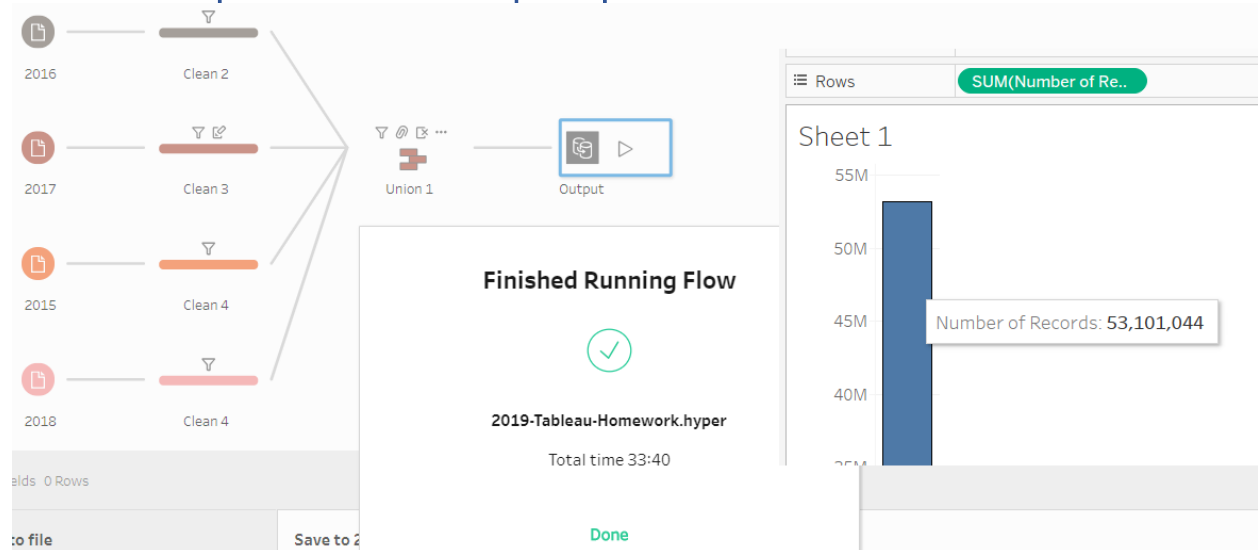
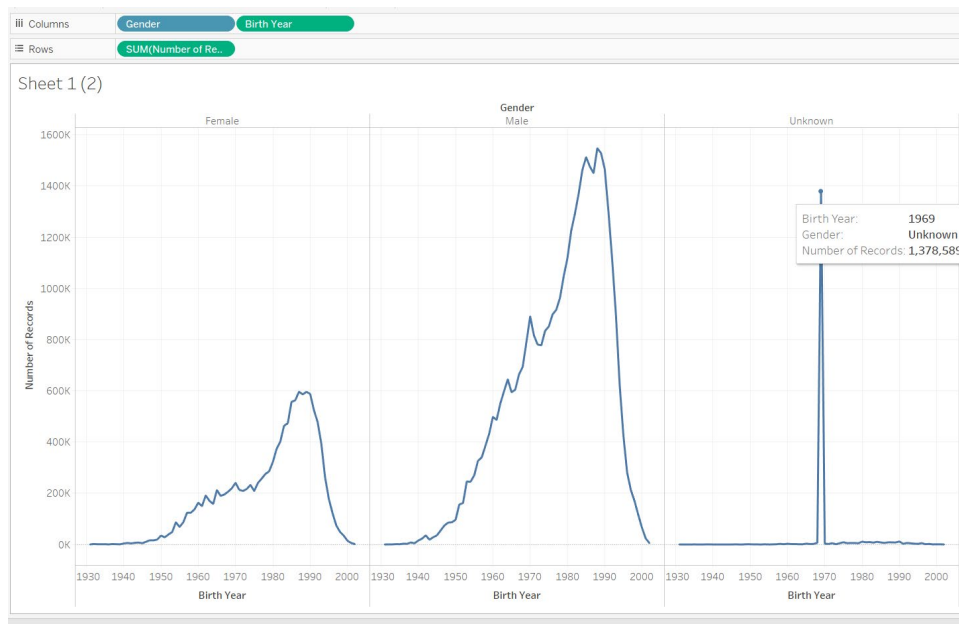


Tableau Prep Builder 2019.3 | Output File.

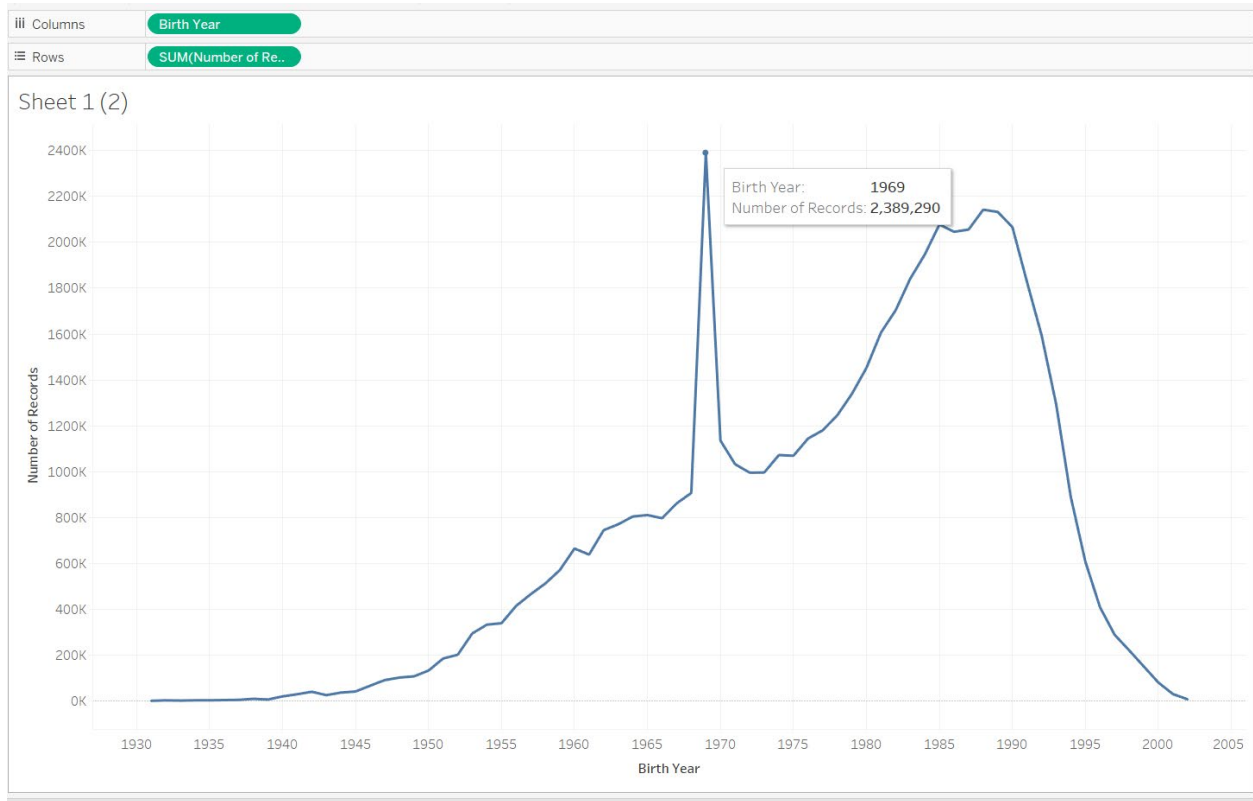


Combining the output of 2015, 2016, 2017 & 2018 with my filters took over 30 minutes to generated providing me with over 53 million records.

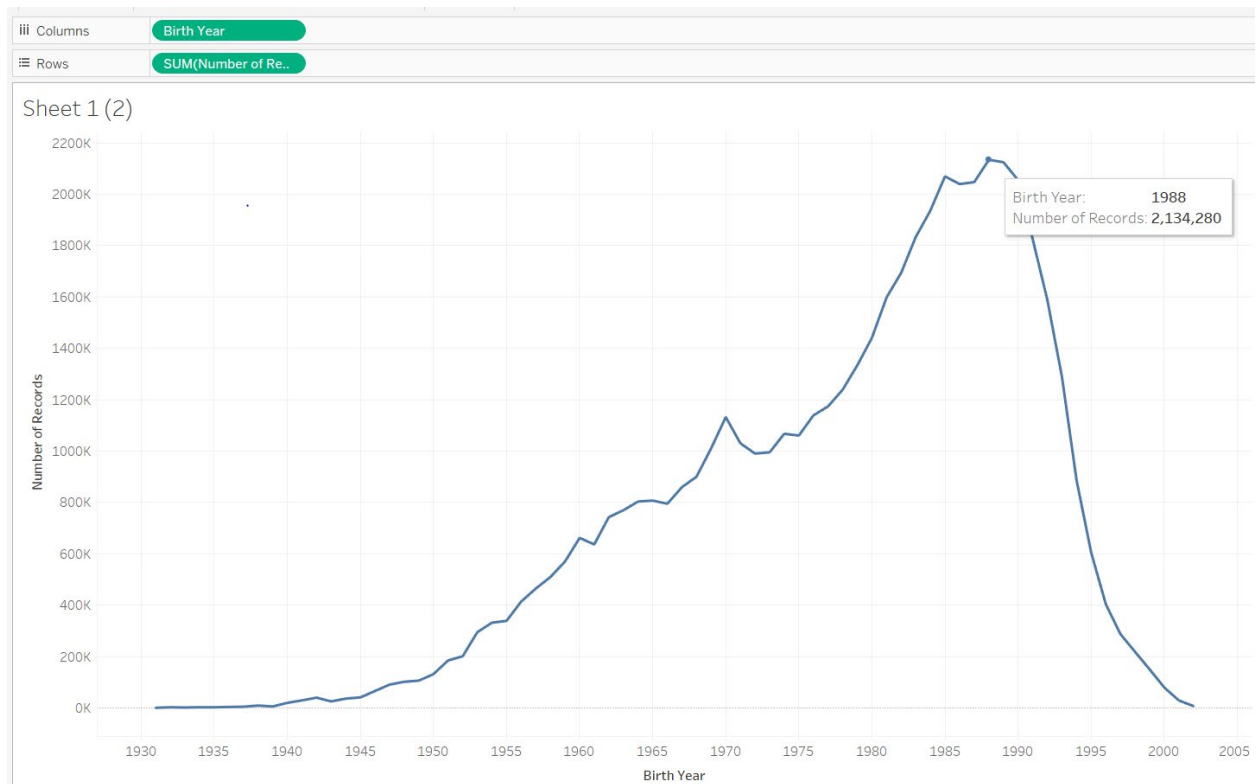
Even more excitement brewed than just with the quick preview from the union stencil output to Tableau Desktop. The 1969 anomaly was even highlighted even more by creating some quick graphs quickly showed a spike in 1969 of 'unknown' gender with over 1.3 million records. I will dismiss this data has probably the unknown defaults in some sort of form entry defaults. Having the Gender, Birth Year as columns and Sum of all records below produced this interesting spike in confused 50 year olds.



By removing the sorted Gender and just using the birth year and sum of records, the spike was quite obvious.



By removing the unknown gender as a filter we achieve a very nice graph.



Data Utilized: 2015, 2016, 2017, 2018:

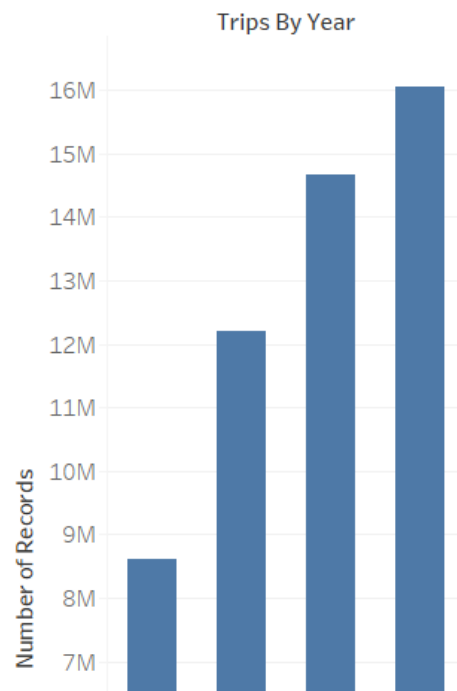
Questions to Answer:

1. **How many trips have been recorded total during the chosen period?**

51.5 million trips have been recorded utilizing the data from 2015 to the end of 2018.

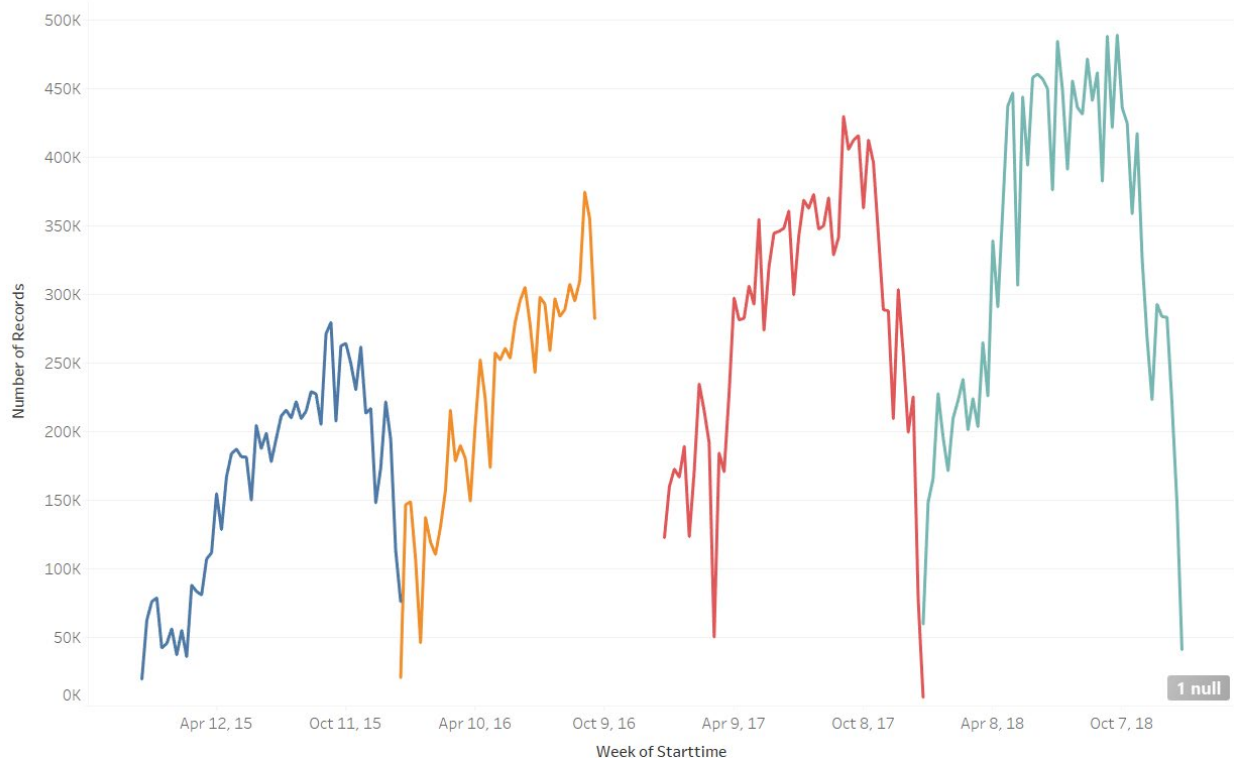
- a. 2015: 8.6 million
- b. 2016: 12.2 million
- c. 2017: 14.7 million
- d. 2018: 16 million

Total Trips by Year



2. By what percentage has total ridership grown?

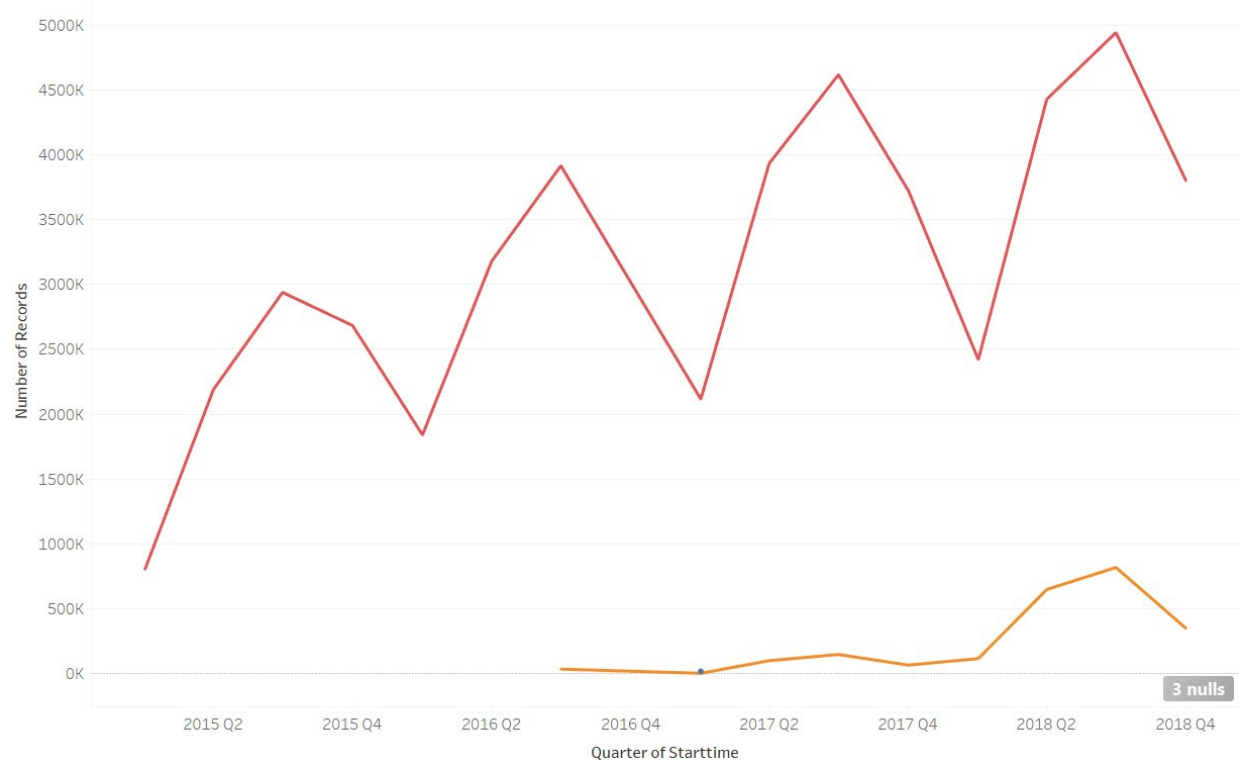
2. Ridership Growth



This rider growth graph shows the growth continuation from Jan 2015 to Dec 31th 2018. There is a gap in the data for Q4 of 2016 between September 26th 2016 to Dec 31st 2016.

3. How has the proportion of short-term customers and annual subscribers changed?

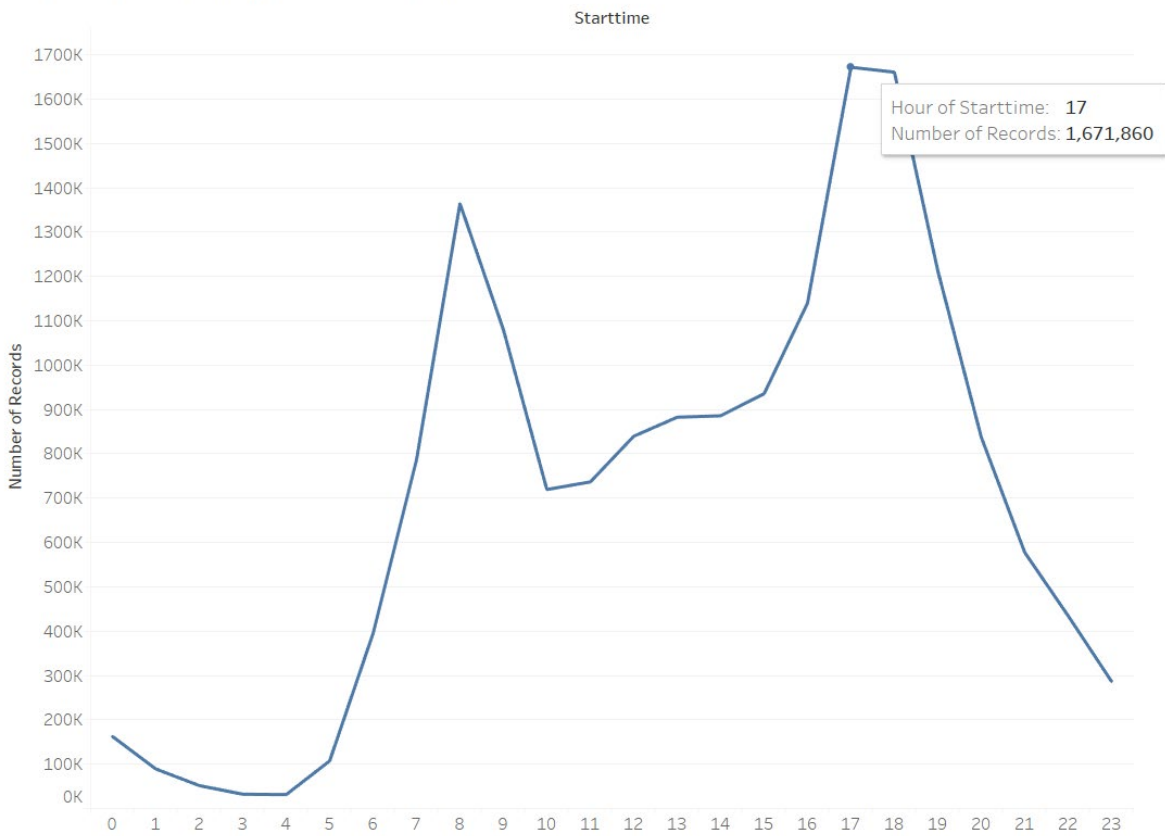
3. Proportion Customers Vs Subscribers



It's logical to see how the drop off of subscribers in Q1 of every year it being the winter months in New York.

4. What are the peak hours in which bikes are used during summer months?

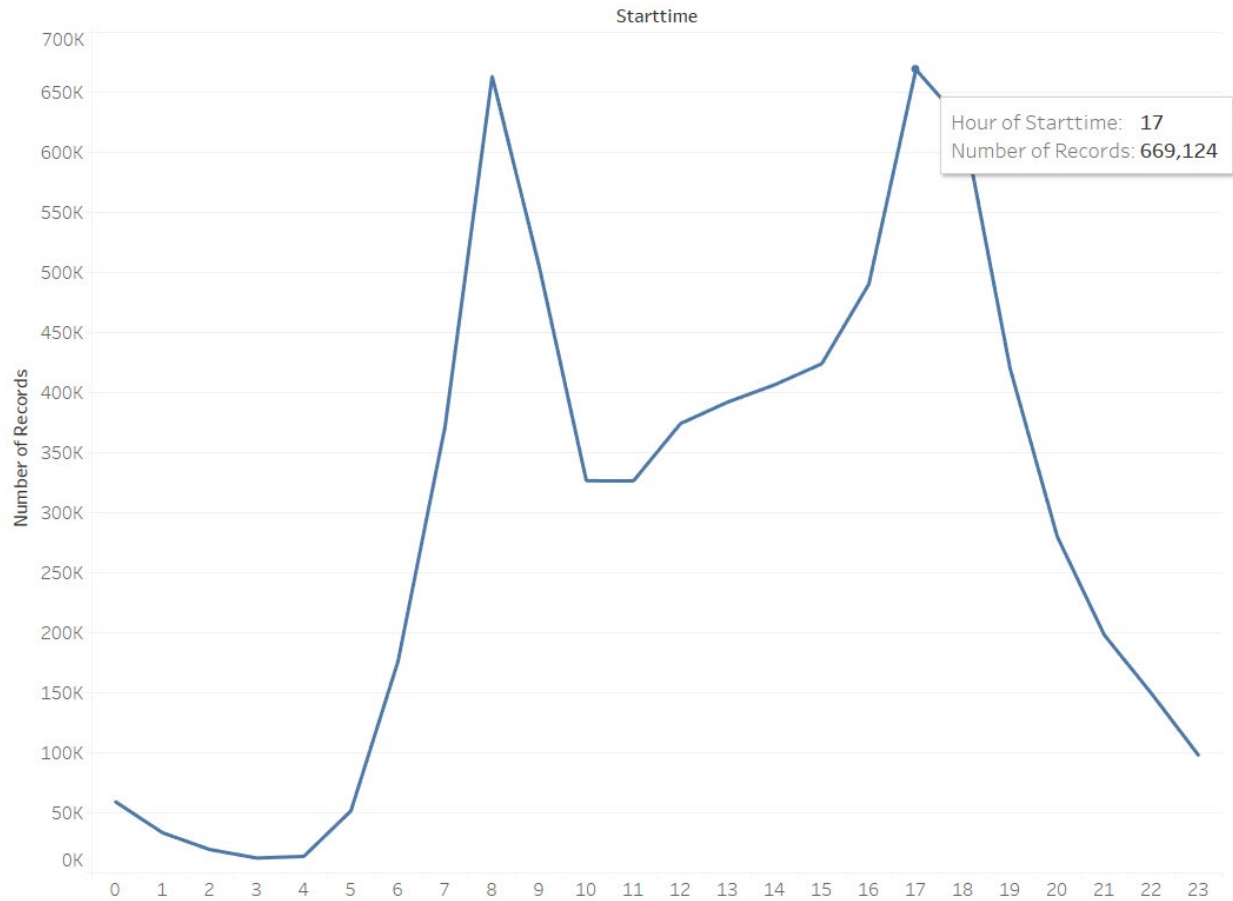
4. Peak Hours During Summer Months



Again no surprise here, very logical that the peak hours during the summer months of June, July & August are early morning at 8am and peak between 5 & 6pm.

5. What are the peak hours in which bikes are used during winter months?

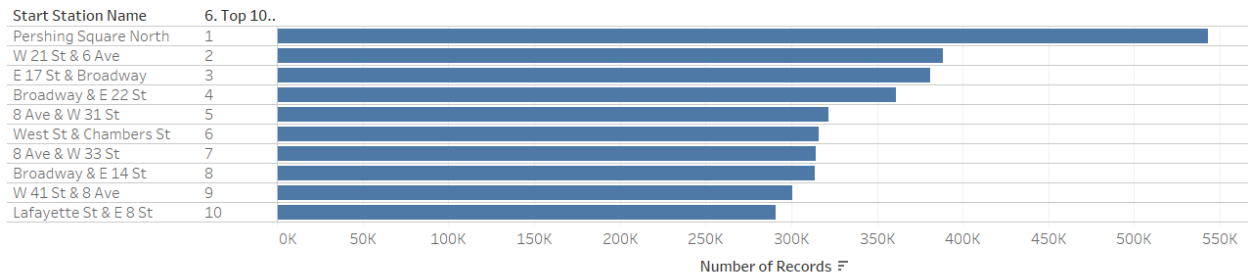
5. Peak Hours During Winter Months



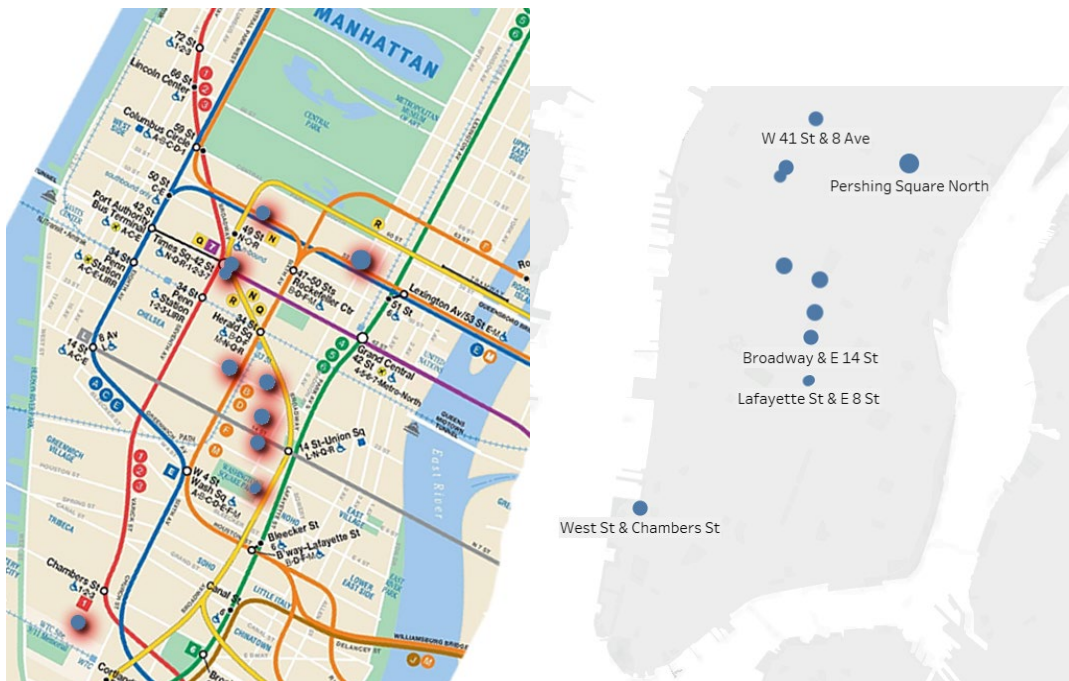
Again no surprise here, very logical that the peak hours during the summer months of December, January & February are early morning at 8am and peak at 5pm.

6. Today, what are the top 10 stations in the city for starting a journey? (Based on data, why do you hypothesize these are the top locations?)

6. Top 10 Starting Stations

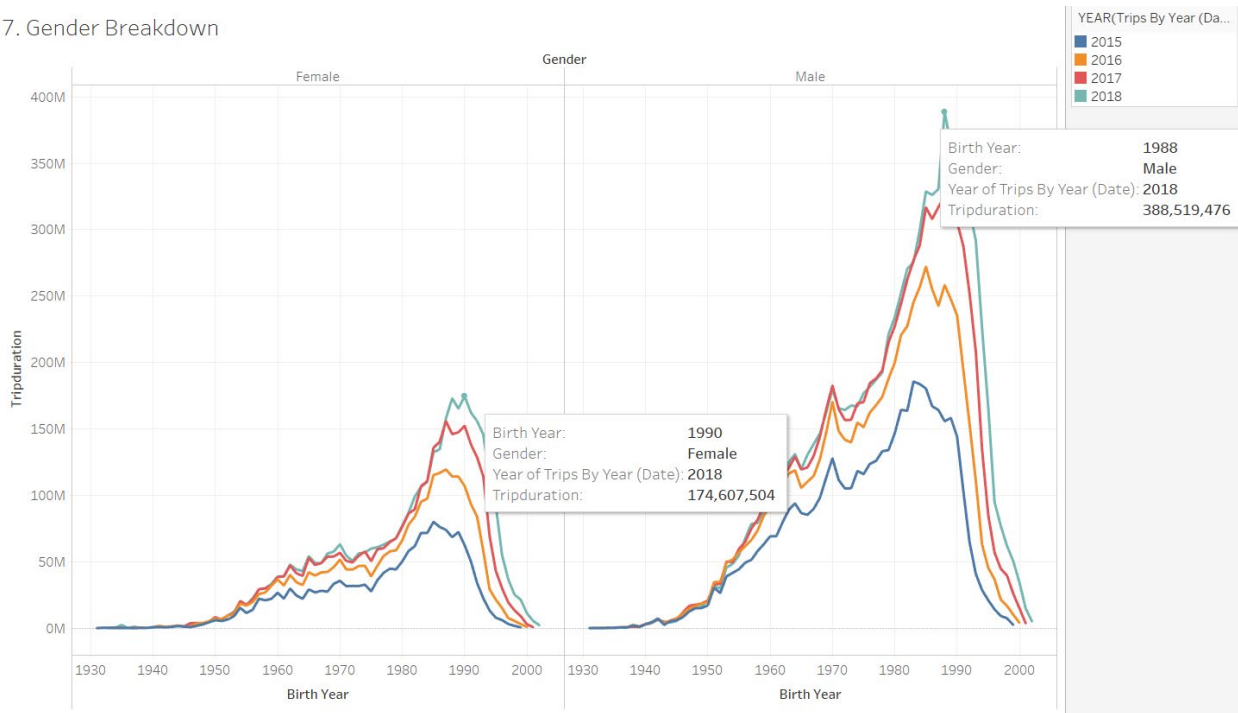


So it seems that the top 10 starting stations are relatively close to New York City Subway line.



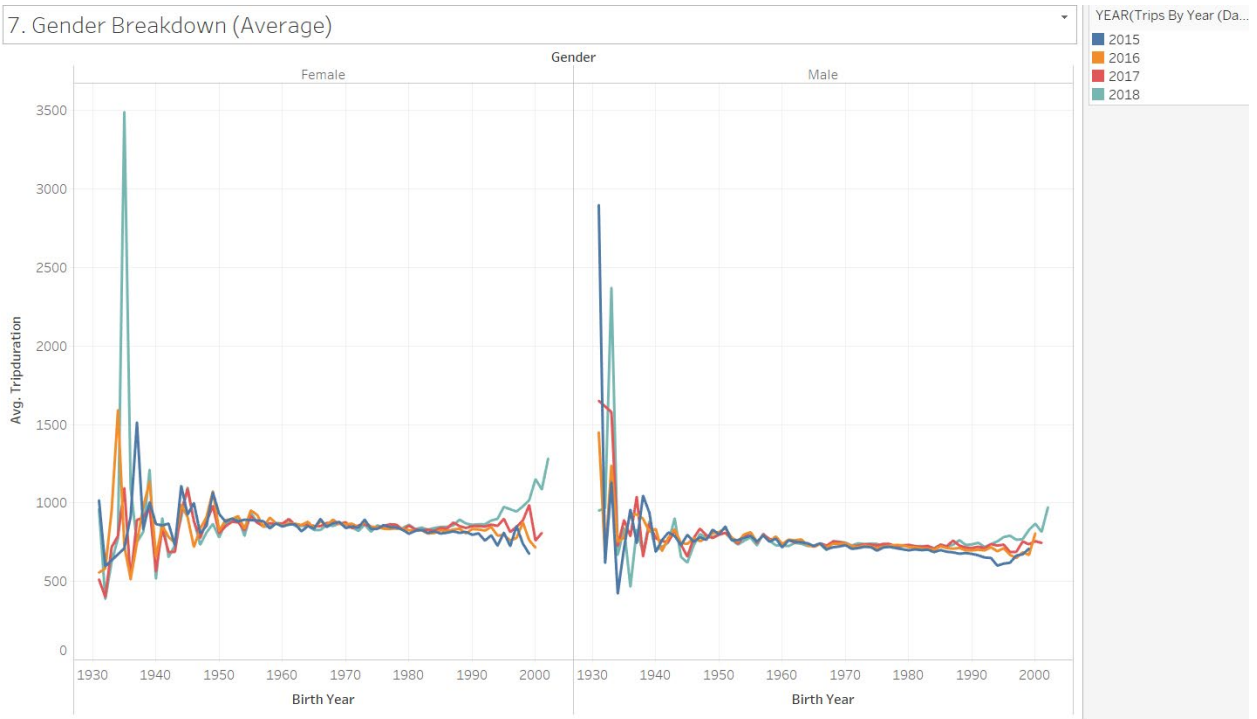
7. Today, what is the gender breakdown of active participants (Male v. Female)?

7. Gender Breakdown



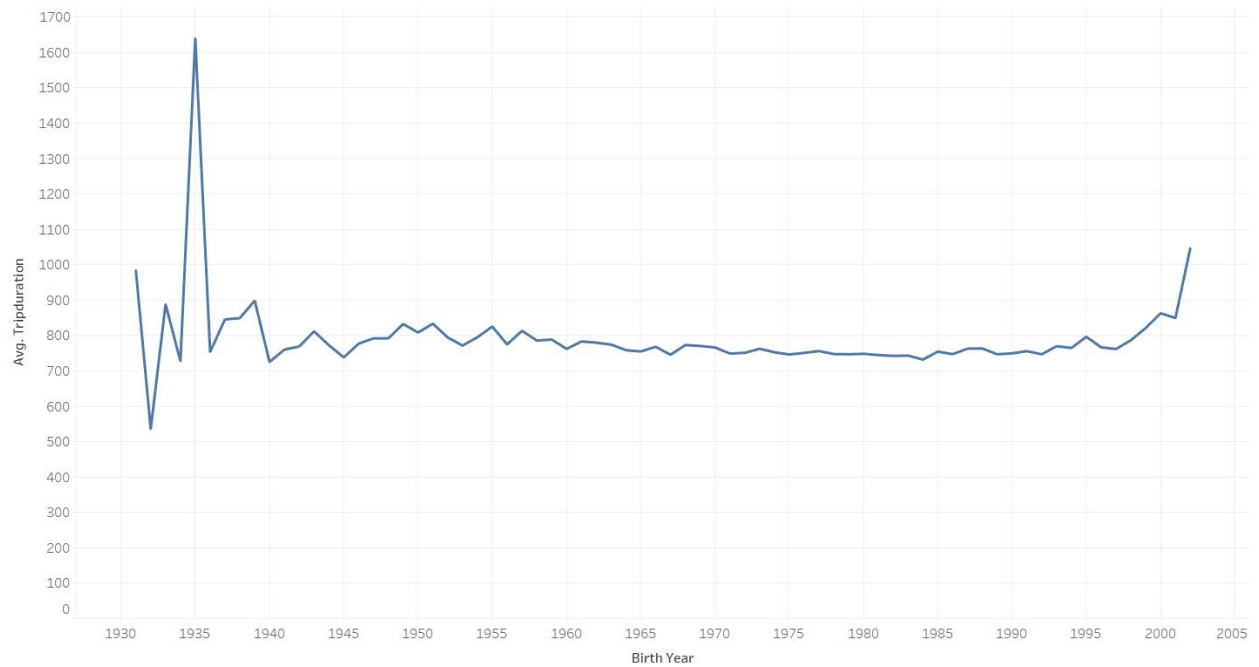
Here is a chart that shows gender by birth year with the **Sum** of the duration of the trips. While a vastly different graph below shows the **AVG trip** duration is much more weighted towards older subscribers enjoying the old style of transportation, perhaps during their bike through central park during the day. I believe strongly that the 1935 / 2018 data for both male and females need to be weeded out. There a smaller spike for males born in 1931 for the 2015 data set.

7. Gender Breakdown (Average)



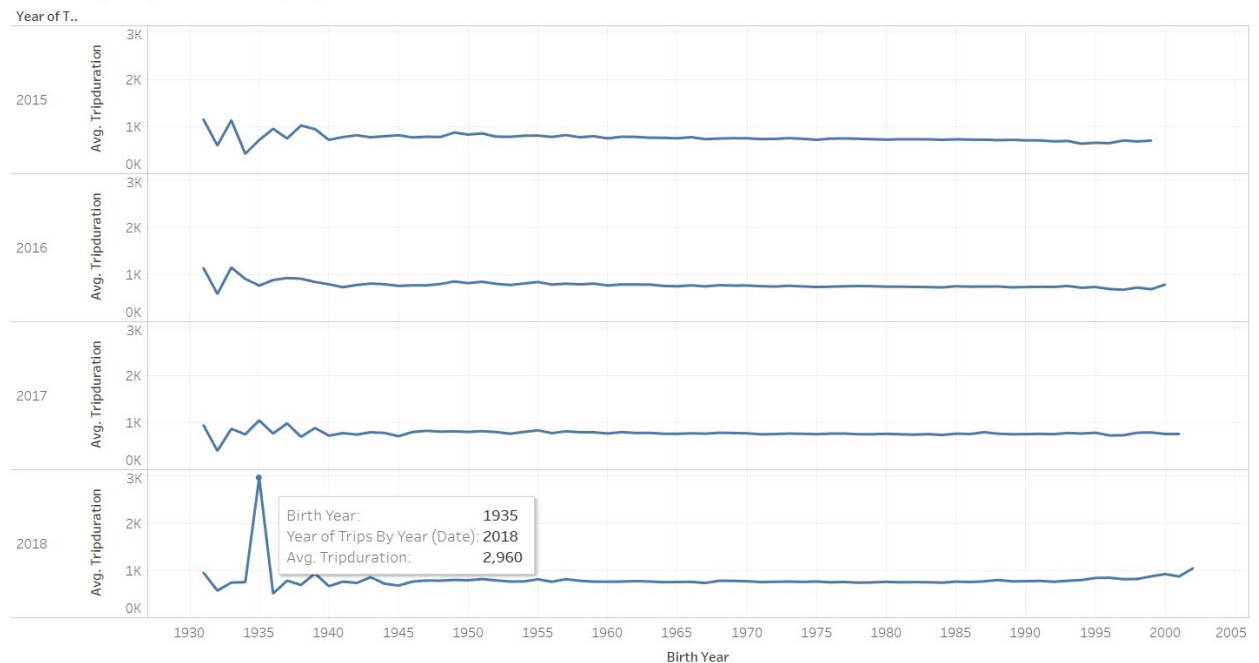
8. How does the average trip duration change by age?

8. Average Trip Duration By Age



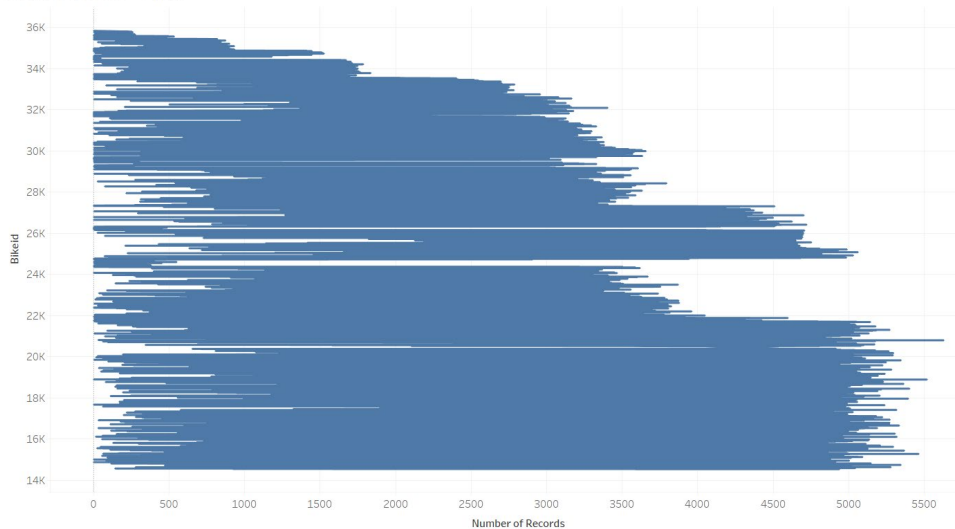
It would be worth diving deeper to analyze the spike in people born in 1935. After parsing out the graph, it's easy to see how there's an anomaly in the 2018 data which is the root cause of the odd number spike.

8. Average Trip Duration By Age



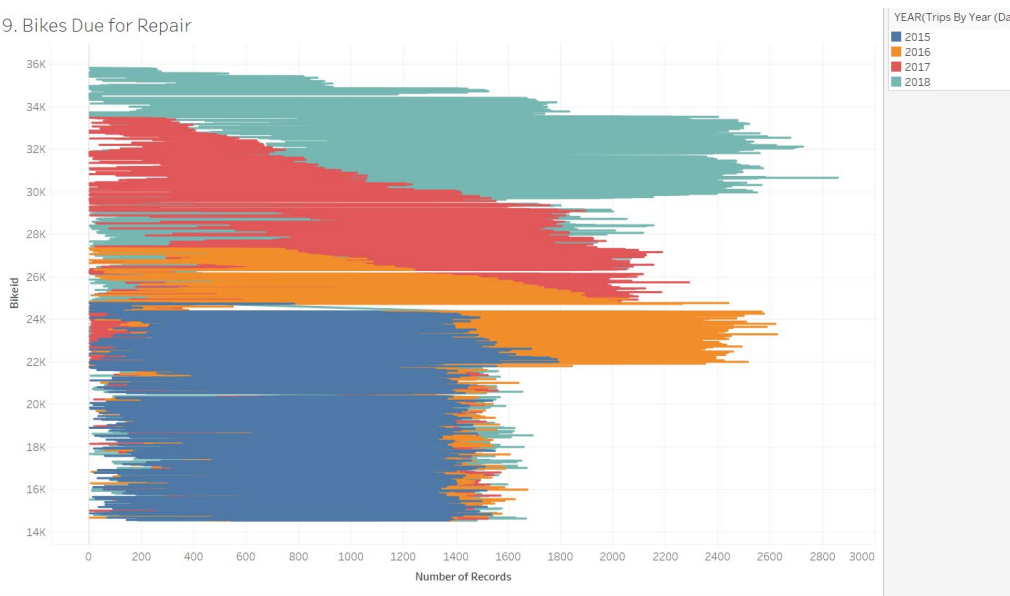
9. Which bikes (by ID) are most likely due for repair or inspection in the timespan?

9. Bikes Due for Repair

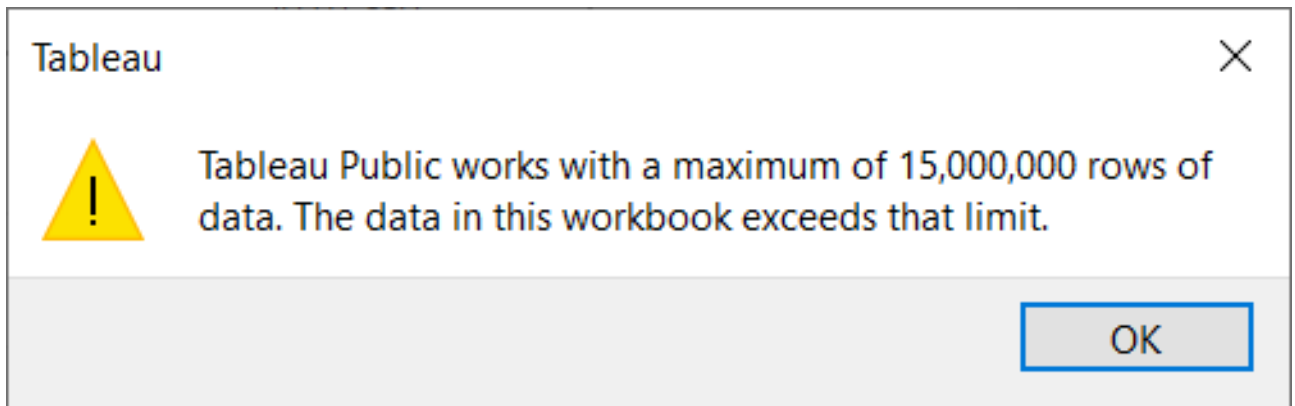


While playing with the charts an interesting point revealed itself. The above chart is showing the total records by Bike ID and it's logical that the lower number bike ids (in the 16k to 18k ranges) are the older bikes that have been on the services the longest since 2015 which begged me to ask the question how could I show that which brought to me to this graph by adding the year as a colour dimension.










9. Bikes Due for Repair



There seems to be a real explosion of new bikes brought into service in 2016 and then a whole new wave in 2018.



Well isn't that just delightful...

Name	Date modified	Type	Size
 2015.csv	11/3/2019 8:48 PM	Microsoft Excel Comma ...	1,818,041 KB
 2016.csv	11/3/2019 6:55 PM	Microsoft Excel Comma ...	2,547,480 KB
 2017.csv	11/3/2019 6:56 PM	Microsoft Excel Comma ...	2,836,824 KB
 2018.csv	11/3/2019 6:57 PM	Microsoft Excel Comma ...	3,258,515 KB
 2019-Tableau.tfl	11/15/2019 3:32 PM	Tableau Flow File	10 KB
 2019-Tableau-Homework.hyper	11/15/2019 3:31 PM	Tableau Extract	1,578,304 KB
 2019-Tableau-Homework.tde	11/15/2019 2:47 PM	Tableau Extract	30 KB
 2019-Tableau-Homework-Desktop-Workbook.twb	11/15/2019 8:28 PM	Tableau Workbook	1,263 KB
 2019-Tableau-Homework-Original.hyper	11/15/2019 3:31 PM	Tableau Extract	1,578,304 KB