IAT 814 - Visualisation and Visual Analytics

Final Project

Subway Delay Data Trends

Mohamed Anan Anaikar (301428793)

Kritu Patel (301469200)

Introduction

From 2014 to 2022, there were more than 800,000 delays recorded on the subway, Scarborough RT, bus and streetcar, adding up to nearly 13 million lost minutes[1]. These delays constitute a daily factor that impedes on everyday lives of commuters. However historical data showcases that nearly 66% of these delays could have been avoided by proper planning and preparation as the delays often follow a trend pattern [2]. With nearly 50% of Torontonians using the Toronto Transit Commission (TTC) at least once a week there exists a need to solve this issue. However to solve this problem would lie in its core to improve the planning and preparing for the situations that arise and making preparations to combat these challenges. The responsibility to do so lies with the planning and organizing committee of the Toronto Transit Commission.

This project aims to aid the organizing committee of the TTC in understanding the trends in delays across the subway lines in the TTC across months, days and hours from the period of 2014 to 2022. It consists of information such as the duration of delay, its cause, site of incident and when the incident occurred. Using this information the organizing committee will be better equipped to handle the potential delays by preparing ahead for it.

To Summarize:

- Project aims to provide delay trends over Stations, time of the day, causes and months
- Our audience is the TTC organizing committee and operators to aid in improved and on-time transit planning

Problem

Some of the questions that our tool will be able to answer are:

High-Level questions

- Which Stations have the most delays?
- What are the most common causes of delays?
- What is the number of delays that occurred over the last few years?
- When are the delays most frequent?
- Which season has the most delays?
- Which subway lines have the most delays?

Dependent Analysis:

- At a particular hour which are the most causes?
- For a particular cause when and where is it most prevalent?
- For a selected station, what are the most frequent causes and when are they happening?

Data Description

We required data from multiple sources with the primary data available to build the dashboard. We used GIS data to develop maps to develop maps, GIS data of the subway stations data from Google Maps[4]. For the data our primary source was the Open Toronto site which consisted of all the delay related information. We then needed to match the station location and name which we gathered from the TTC site[5] to access and develop the lines of the subway.

Document name	Source	Link
Subway Delay	Open Toronto	https://open.toronto.ca/dataset/ttc-subway-delay-data/
Station	TTC site	https://www.ttc.ca/routes-and-schedules
Map	Open Toronto	https://open.toronto.ca/dataset/ttc-subway-shapefiles/

In terms of dimensions the data that we are using consists primarily of

- Delay in mins
- Station Name
- Code (Causes)
- Date
- Stations

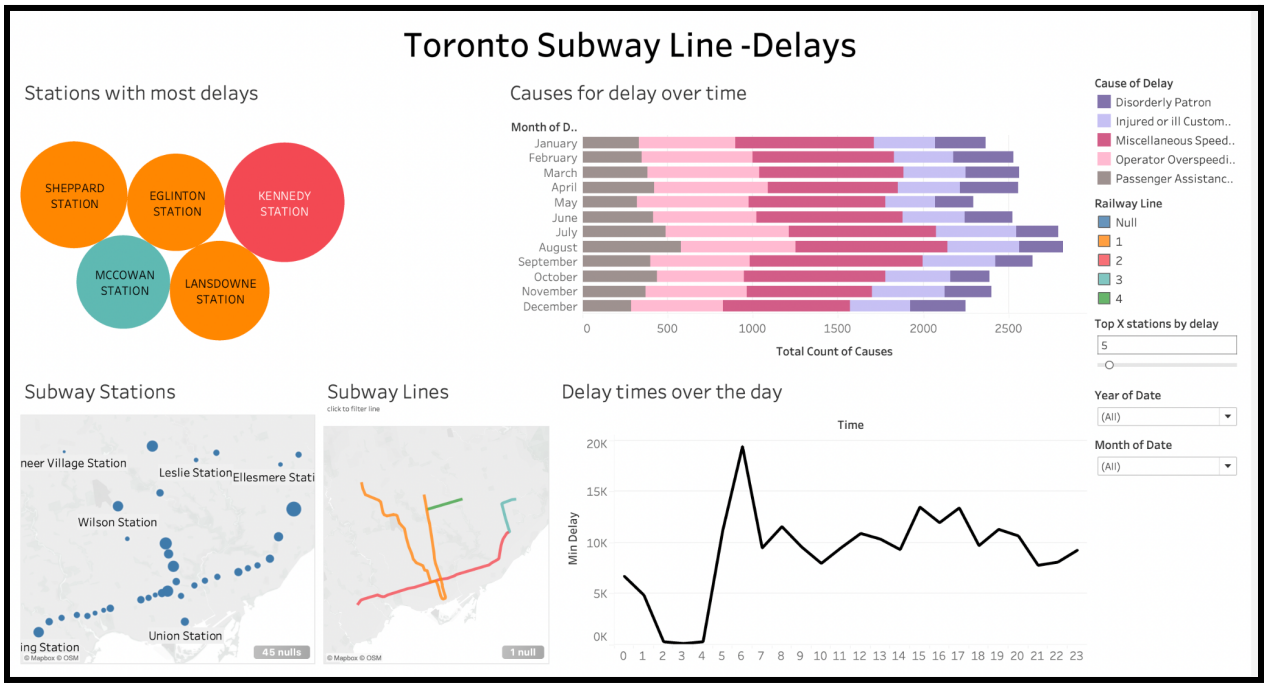
Implementation

The delay data had been divided annually from 2014 to 2019. Our first task was merging the data to understand the outlook and identify trends over the years and seasons. We wrote a python script to merge the excel files for all the years matching the columns and dropping fields that were not common between the years. We then cleaned the “dirty data” i.e. the fields that were lacking the key dimensions we have prefaced above. The data cleaning process was done using Tableau prep to handle the subset generated from Excel as Tableau Prep initially could not handle the amount of data. The task that was most difficult finally was forming the tables in tableau for the aforementioned three datasets (GIS, Stations and Delays) in Tableau as the order was the most important to enable creating the mapped data. We first formed a left joint with Delays and Stations and full joint to the union before this for the GIS information.

5 Visualizations

To aid in answering the questions that were presented above in the problems section, we have developed two dashboards - first serving to provide a holistic view of delays across stations, lines and causes that has been curated by aggregating; second providing the ability to compare between causes across stations and times of the day.

Dashboard 1: Delays Across Toronto



The above visualization aids in the understanding of the High- level questions that were stated, the dashboard is divided into five panes with each pane assisting in answering questions with respect to delays both holistically and with respect to each attribute enabled through interactions.

The bubble chart shows the stations that have the most delays from 2014 to 2021, a key component in deciding the allotment of resources across the stations.

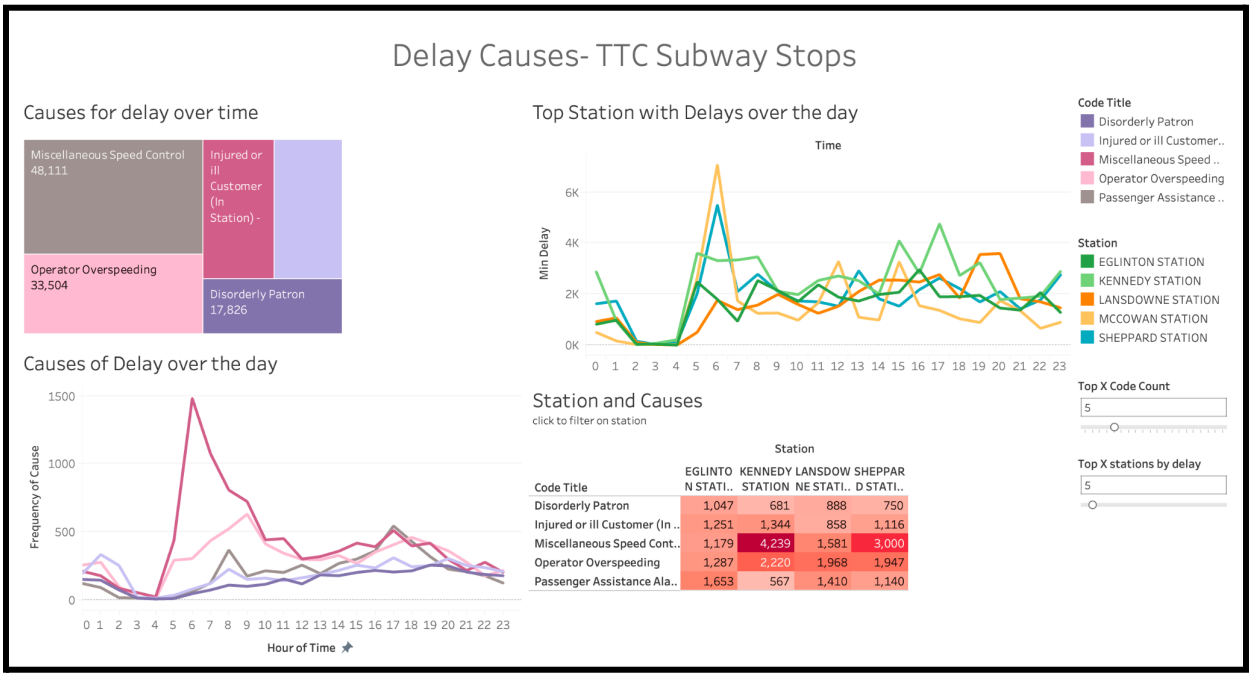
The stacked bar chart pane showcases the most prominent causes of the delay, with the leading 5 causes shown in the stack by months which help in setting up suitable countermeasures for the causes.

Subway Stations consist of all the stations mapped out in the TTC railway line, these points have been sized to increase in point size with the stations having a larger number of causes indicating the level of attention needed for the corresponding station.

Subway lines are the four lines of Railways that have been considered to help better understand trends and filter for a particular line of information to understand the delays across a particular route better.

Delay Times over the day showcases the total count of delays occurring at each particular time to enable personnel assignments better throughout the day.

Essentially nearly all the high level questions could be answered by just each section of the above dashboard where any individual aggregation can serve as an answer for delays across stations, time and causes



For the second Dashboard aims to answer questions on a finer level, as opposed to a generalized approach in the previous dashboard this dashboard will provide the decision makers with greater specificity to help combat various issues that arise in the railways.

The tree graph of causes over delays showcases the various causes of delay over time with an associated label to understand the causes better and make the suitable adjustment for the causes.

The two line graphs help answer the key questions associated with Time

- What is the Cause of Delay
- Where is the Delay happening(Stations)

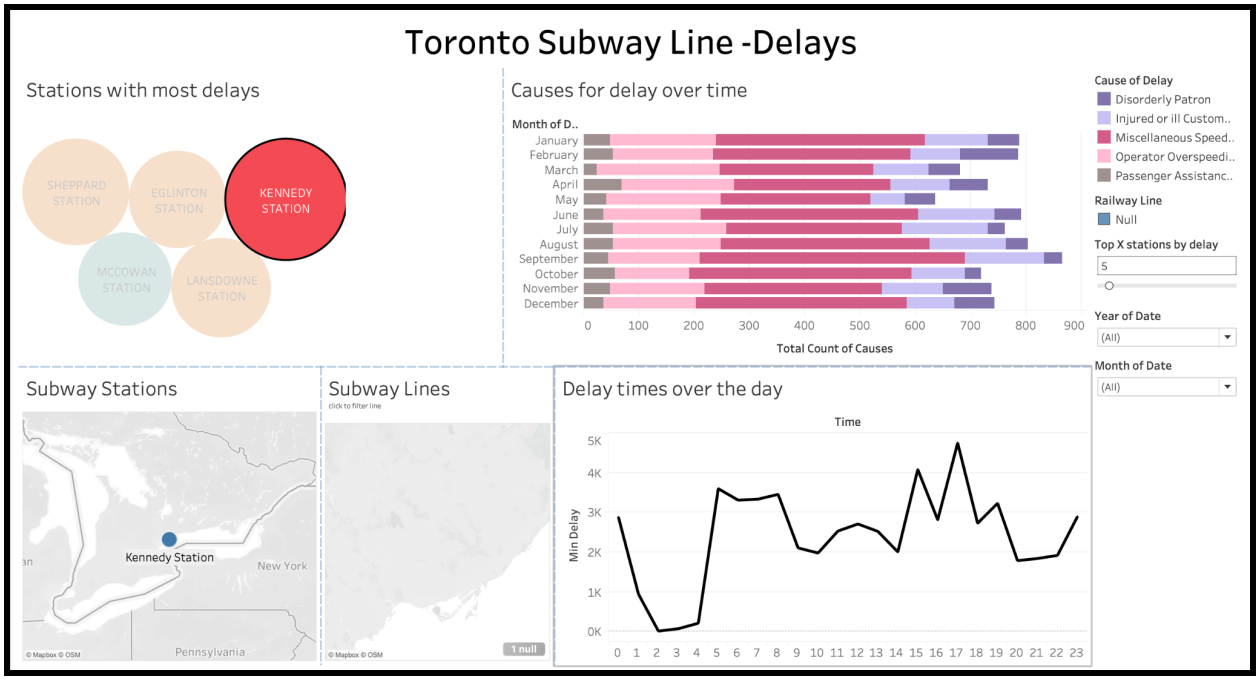
This finer level of refinement makes the personnel allotment decisions more concise and apt for each location and time.

Finally the heatmap aims to bring to notice the greatest causes with location in a more clear manner with respect to stations while offering a good comparison of other causes in the station as well.

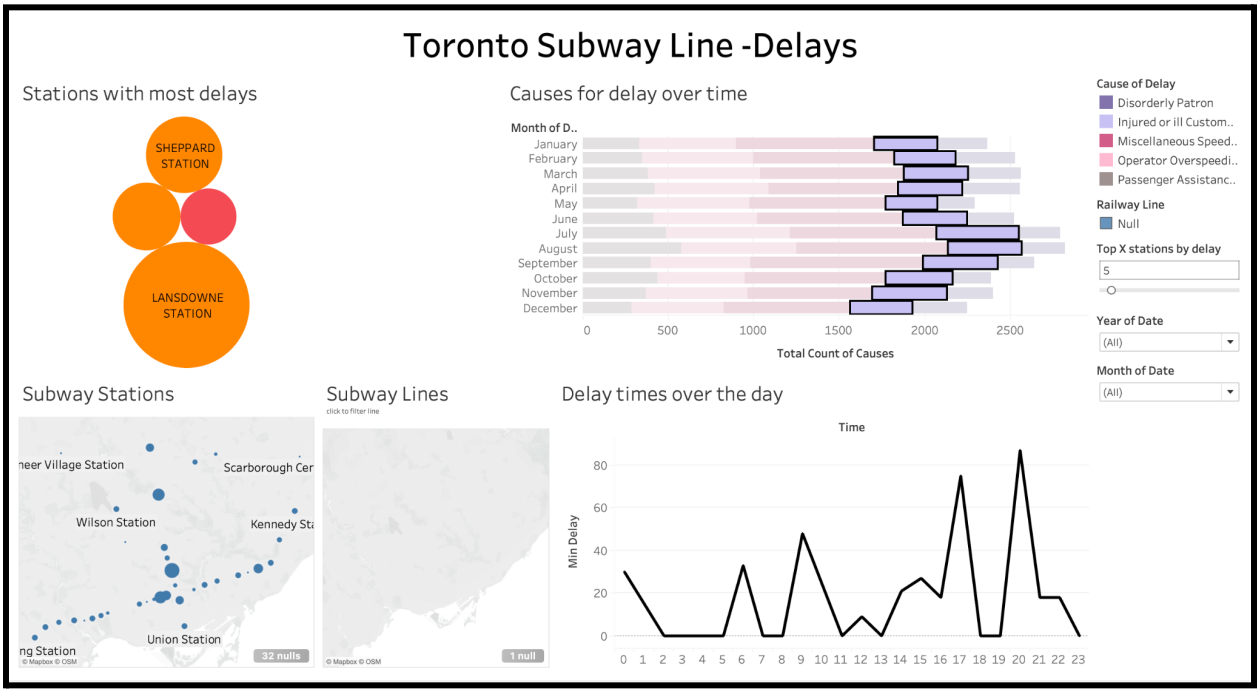
Interactions

A key component of our dashboards is its interactivity. All the sections of the dashboard can be used to further filter down and answer key questions such as the time of day with most delays for a given station.

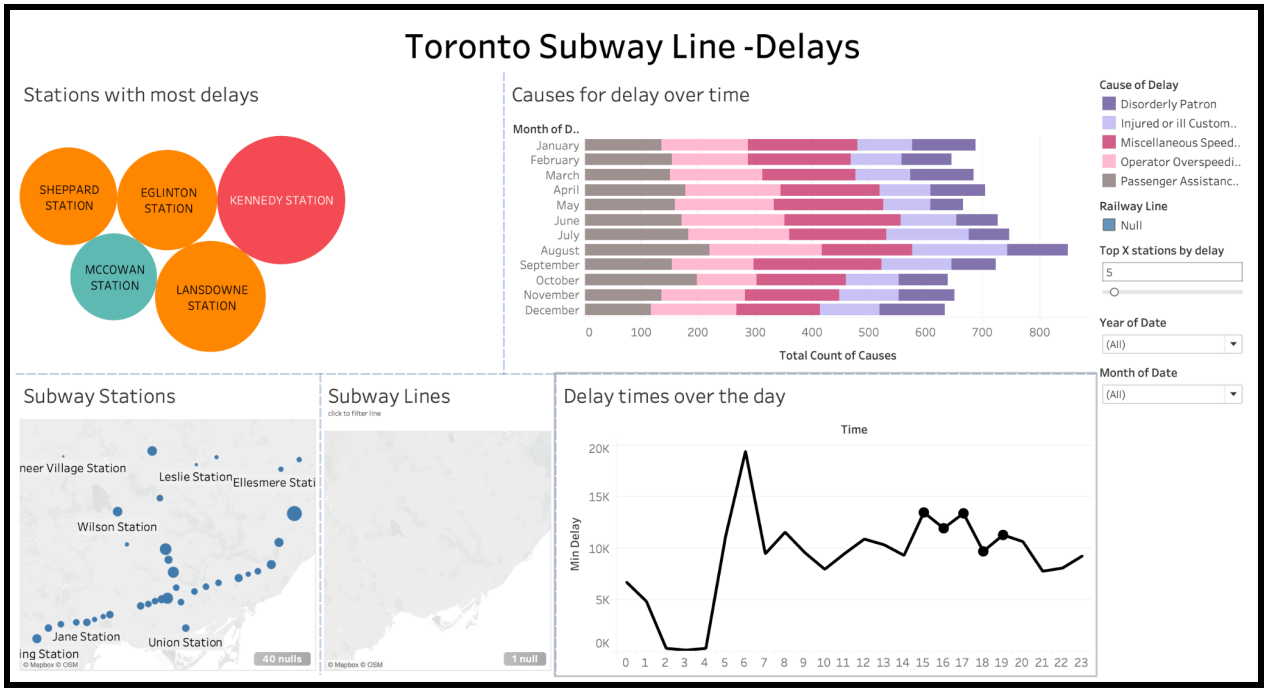
As for the above example, on selection Kennedy Station in the Station with most delays section the corresponding sections all shift to showcase all the attributes corresponding to the station selection which now shows that Miscellaneous speeding is the leading cause for delays at Kennedy station with 6 pm being the peak of these delays.



Now conversely with the cause as our main attribute, to see where and when most injuries or ill customers are appearing, we get the visualization below where we see the stations, their locations and their times - in this case late at night as the peaks for injuries.

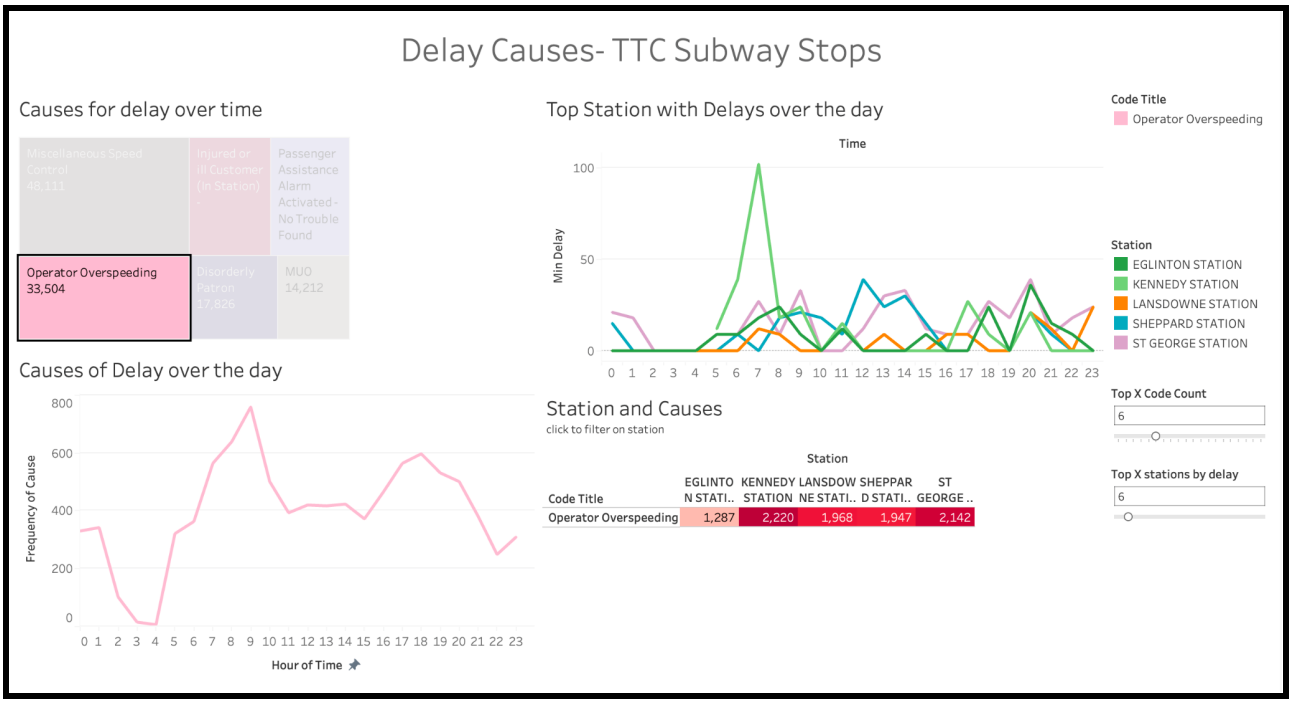


Similarly with time as a primary factor we see the associated aggregations during the period of time from 3-6 pm that Operator overspeeding is the most common cause and Kennedy station having a lot of delays in this time frame. The subway station section also shows all the stations having delays at the particular time.



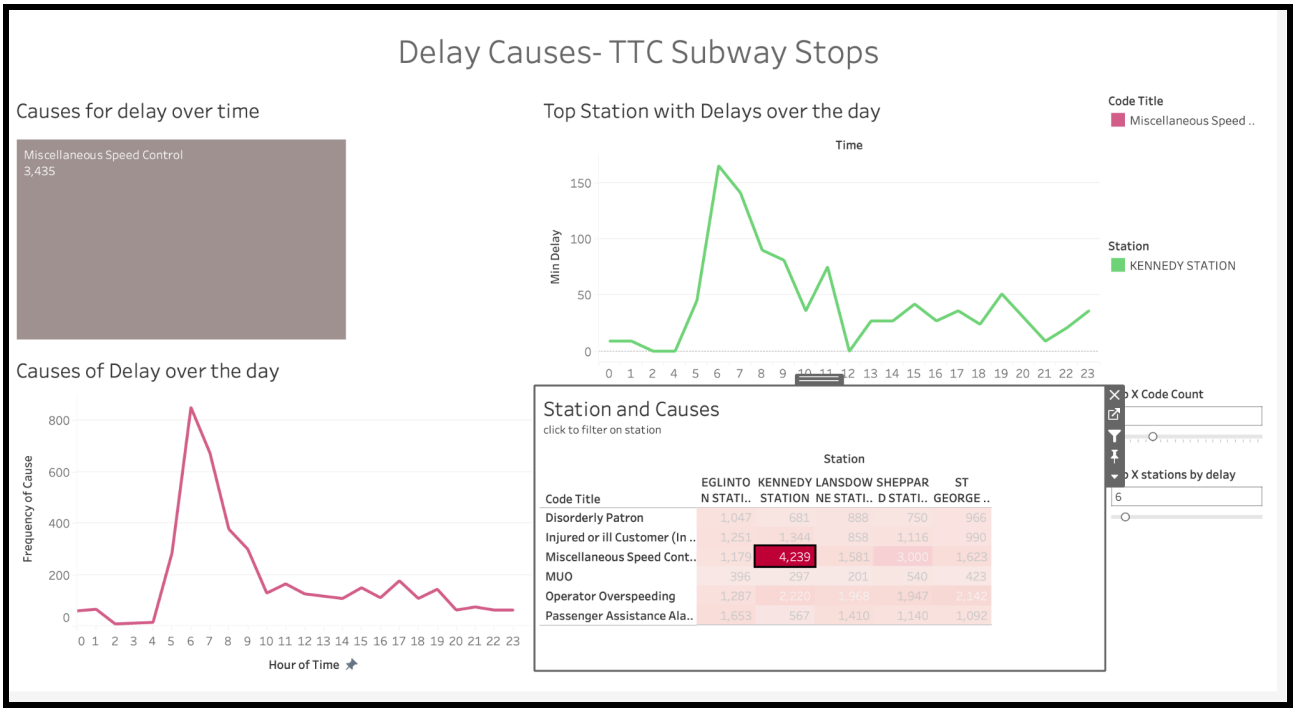
Consequently the subway stations and subway lines can also be used as the focal question points the data aggregates to show the causes and time trends for that particular station or lines.

The second dashboard as stated in the previous section comparison of trends across attributes becomes clearer. For example with operating speeding as the primary attribute selected in the tree chart, questions such as time of the day - 8 am and Eglinton station at 7 am can be seen as the chief fields.



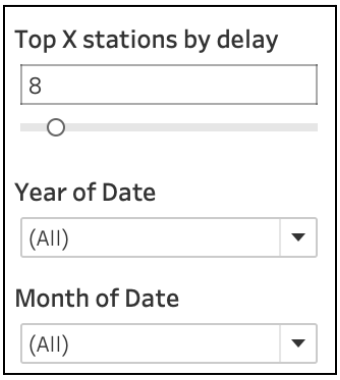
Likewise with a particular time slot (7-11am) and select causes the corresponding station delay counts

Particular selections to visualize trends for that selection can be made from the heatmap to better understand and prepare for the intersecting values for example Kennedy Station's Miscellaneous operator overspeeding

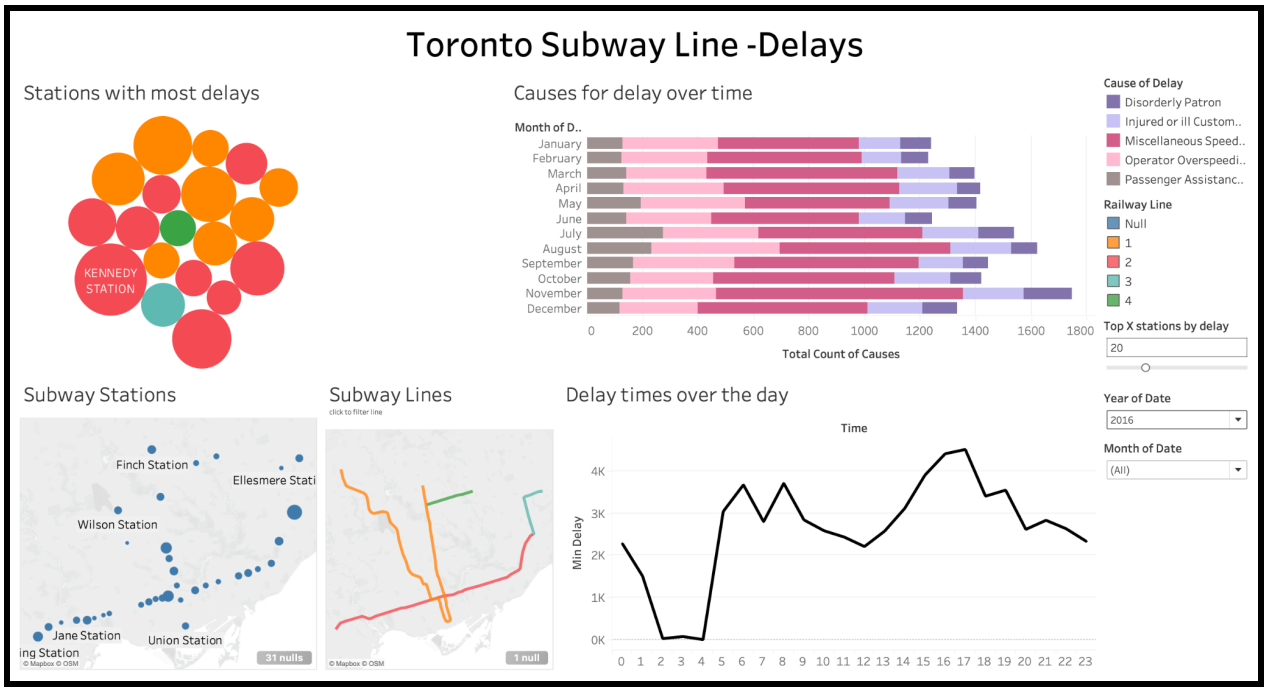


Filters

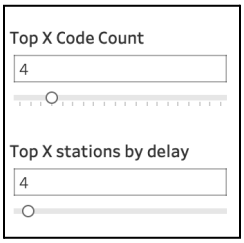
Both dashboards consist of filters in the right side which allows to make the data more accessible and answer more questions as opposed to limiting values in just the sheets. With Top X stations referring to the number of stations you would like to show



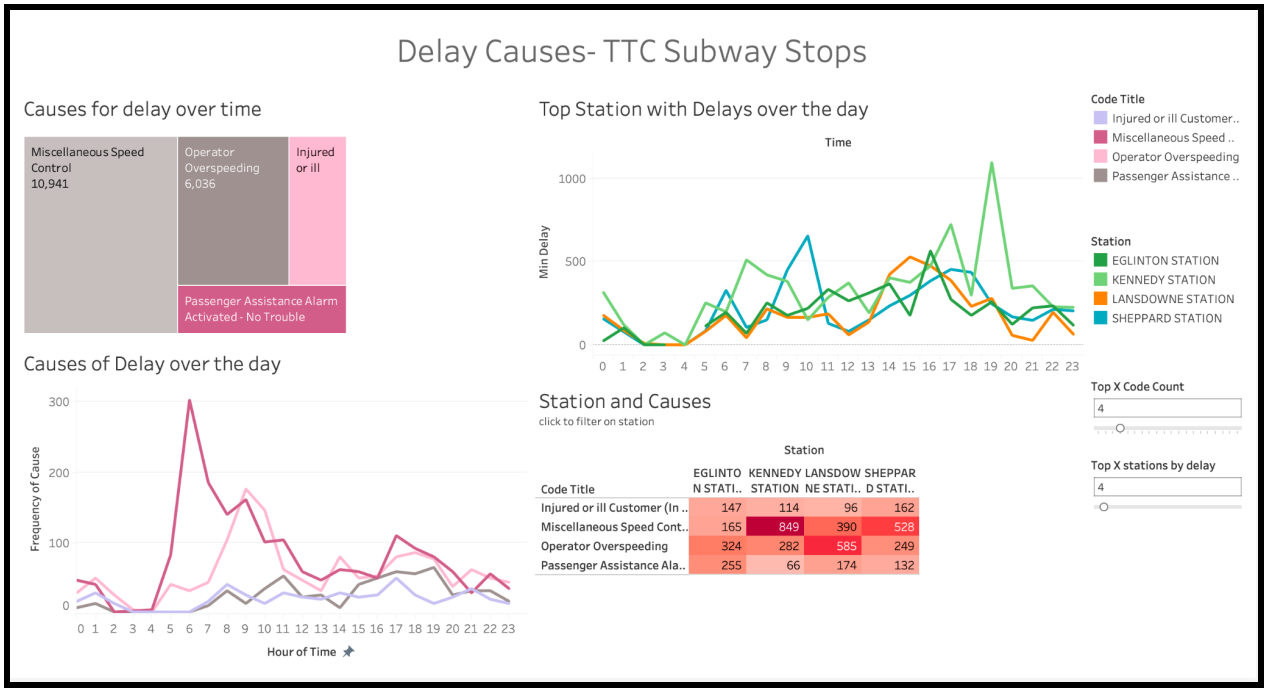
An example below would be increasing the number of stations to top 20 and 2016 significantly changes both the outlook of the map and its corresponding questions it would answer



Similarly for second dashboard, we have the filter to showcase how many stations and causes of delay the user would like to see



With the filter affecting the dashboard as the following visualization below shows



Design Decisions

With Generalization as the theme for the first dashboard and specificity as the theme for the second the corresponding decisions were made to match the aforementioned themes:

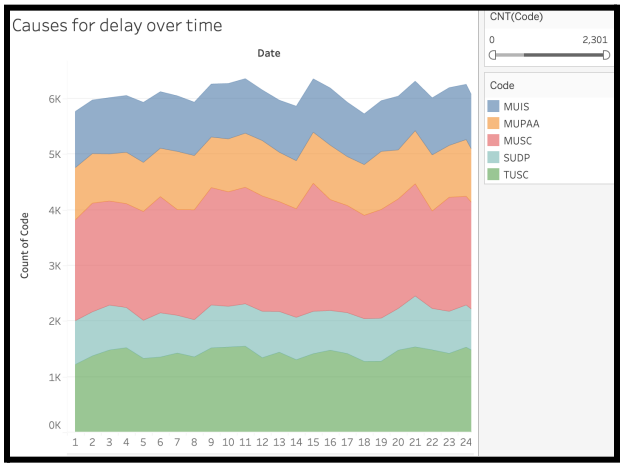
- Line Graphs with time of the day as the centralization attribute as time was considered to showcase differences in the second dashboard
- Hue themes matching across dashboards to enable to users to match a hue with a cause
- Months grouping to show seasonal change values
- Filters to number of Stations
- Red Hue indicating increased number of stops

Improvements

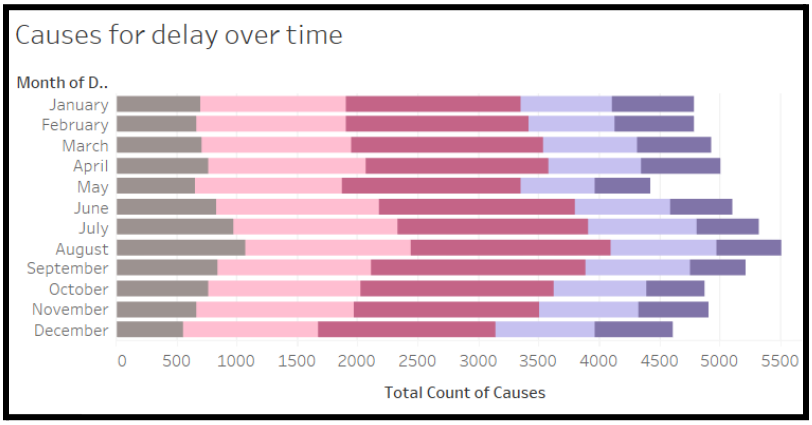
User Feedback :

- Changed the stacked area chart to stacked bar chart as per test user feedback

Before

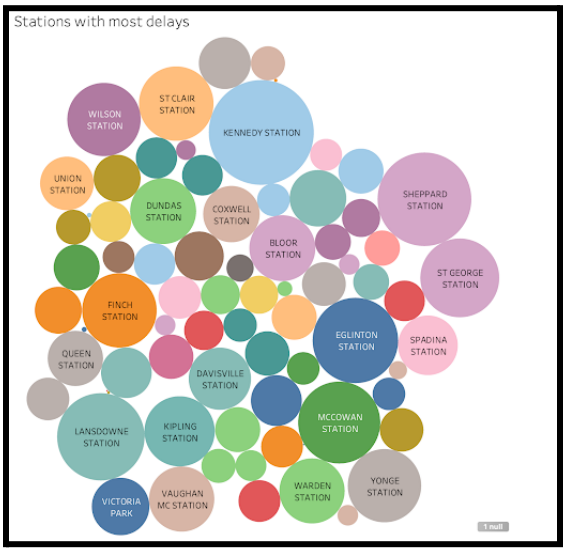


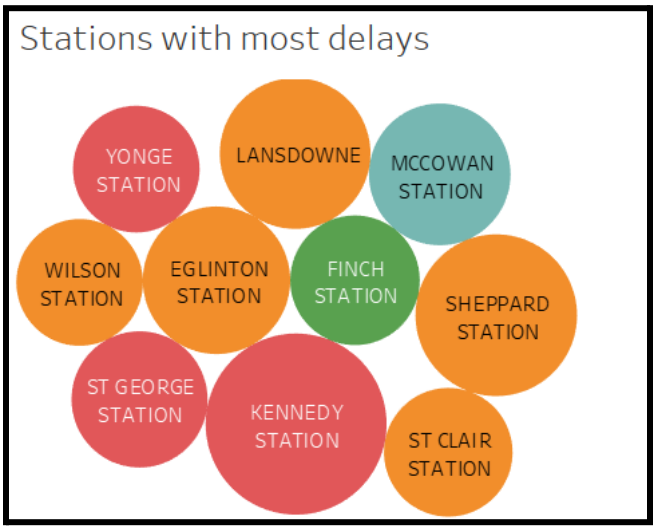
After



- Added filter on top stations to prevent data overload

Before **After**





Future Works

We aim to provide a better interaction between the station subsection in the dashboard as some of the data is incomplete and returns null value, including better interaction on clicking the station and between station to top station value. We were hoping to provide a mechanism to shift between months and years in the new dashboard as currently it is only classified by months. Additionally we would have really liked a better color collection as a better way to associate particular colors to particular stations. Finally we would have really liked to provide a single view for both the dashboards.

Conclusion

The Toronto subway system, being a vital transportation lifeline for millions of commuters and tourists, requires prompt and effective management of incidents and delays to ensure seamless travel experiences.

In conclusion, this dashboard is a crucial tool for the Toronto subway authorities to effectively manage incidents, reduce delays, make data-driven decisions, enhance communication, and optimize resources. By leveraging real-time data and insights, the

authorities can ensure the smooth and efficient operation of the subway system, which is vital for the millions of commuters and visitors who rely on it every day.

References:

[1]https://www.thestar.com/news/gta/2023/03/02/attention-ttc-passengers-the-delay-youre-experiencing-on-the-subway-is-longer-than-ever-heres-why.html?rf

[2]https://www.iheartradio.ca/newstalk-1010/news/poll-public-transit-and-the-ttc-1.3349571

[3]https://www.cbc.ca/news/canada/toronto/ttc-subway-delays-1.4068358