BUSted

A VISUAL ANALYSIS OF PUBLIC BUS ROUTE RELIABILITY WITHIN THE KING COUNTY AREA

IFNX 562

Ana M. Villalpando
Maggie Stark
Sean Miller

08/17/2018

TABLE OF CONTENTS

ABSTRACT

In this paper, we will explore the steps involved in the development and execution of our web based information visualization, BUSted. The goal of BUSted is to help users identify more reliable bus routes for commuting. This paper will cover the steps for producing the Tableau Desktop visualization. It outlines the data acquisition, data cleaning, design prototyping, and iterative user testing processes required to produce our final hi-fidelity prototype. The paper concludes with future work.

INTRODUCTION

The concept behind this project is to develop an information visualization that will illustrate the real-time reliability of predefined schedules and routes within King County to help users decide between routes. Bus commuters rely on applications such as Google Maps and OneBusAway to plan their bus commutes and see how late their bus is. However, when Google presents two or more similarly timed routes the user often must commit to a bus, by going to a stop, without knowing its overall reliability. Our historical look into bus timeliness helps users make an informed decision of which bus to commit to and provides confidence in their on-time arrival.

Target Users and Personas

Our primary users are frequent bus commuters hoping to compare routes options to find the most reliable given their commute parameters. Secondary users, include those interested in the evaluation of bus routes or improving the quality of bus tracking applications. Our first iteration of BUSted focuses on users commuting between Green Lake and Downtown, future versions will allow for broader use.

User Tasks

We identified specific tasks, that users would expect to answer while interacting with our tool. These tasks include:

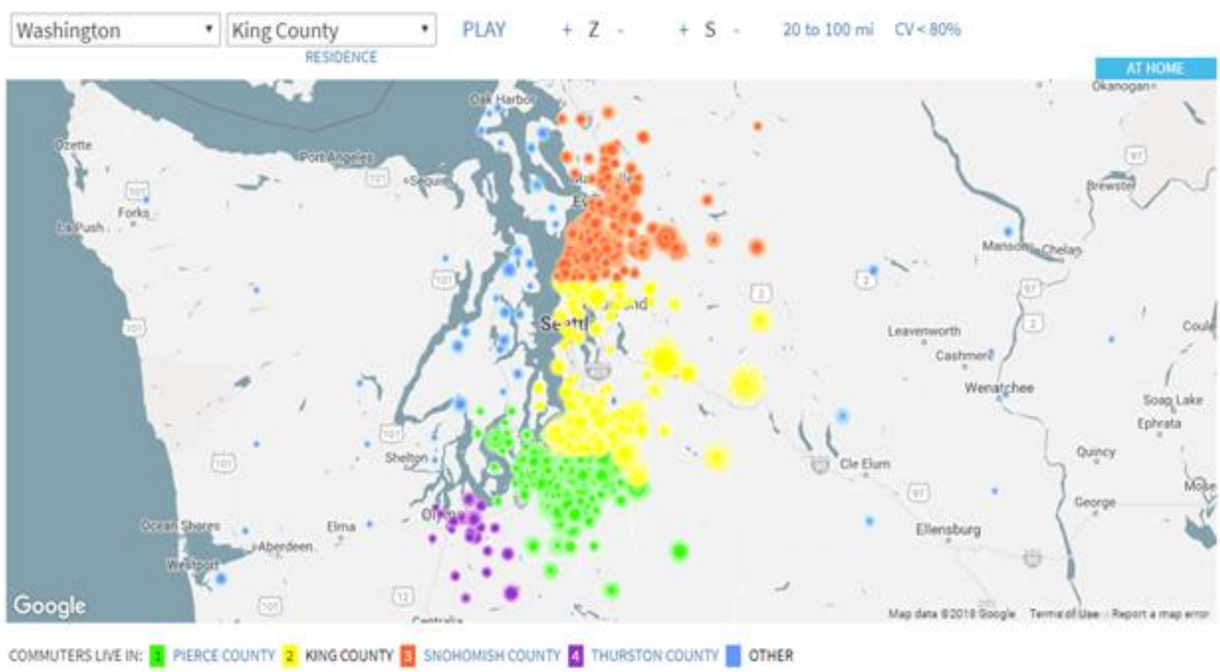
- Which routes are available to me?
- What is the fastest route for my commute?
- Given my parameters, which route is more likely to get me to my destination on time?

Evaluating the effectiveness of these tasks in usability tests helped to refine and improve our visualization.

Previous Work

The Seattle Times published in December 2017 that King County Metro's roughly 175 routes are on-time approximately 77 percent of the time, a metric that has been consistent over the last 5 years.[1] It also explains that King County Metro's definition of on-time ranges from being 1.5 minutes early to 5.5 minutes late, and nearly all routes are late during rush hour.[1] While understanding King County Metro's definition and understanding of reliability can be helpful, their definition of on-time may not meet users' expectations. It also fails to present the data in a consumable format for users to see the granularity of which buses to avoid during rush hour or whether rush hour impacts certain routes more. Our intention was to build upon this analysis with a more user-centered focus, where users can look at reliability on an individual route level.

As far as visualizing commuting is concerned, the available interactive visualizations are either for providing average commute by area (figure 1) or on-demand schedules and delays (figures 2 and 3).[2] The commute map in figure 1 shows the population of commuters given an area. It displays median workplace and residential commute duration by county. While this is effective in showing the volume of commuters by area, where and how the commuters are commuting is obscured into a county wide statistic.



ımmuter flows for workers who work in King County, Washington in 222 unique census tracts, traveling between 20 miles and 100 miles, with C' 80%.

45,386 TOTAL ESTIMATED COMMUTERS ACROSS 1,504 INDIVIDUAL TRACT-TRACT FLOWS

	Rank	County Name	Commuter Workers	Median WCommute	Commuter Residents	Median RCommute	Your Selected Commuters	% This County	% Selected County
	1	Pierce County, WA	318,580	6.3	360,374	7.4	15,742	4.4%	34.7%
	2	King County, WA	1,146,335	7.3	978,552	5.9	14,606	1.5%	32.2%

Figure 1: ACS Commute Map

Two staple applications of a commuter in King County are Google Maps and OneBusAway. OneBusAway is good at allowing a user to see how late their bus is at a given stop while Google Maps is great for suggesting routes, schedules, providing on-demand lateness, and scheduled route durations. Google Maps' interface is designed for even the most novice bus commuter, while OneBusAway requires users to be more familiar with specific routes and their stops. Our design will build upon these helpful features while providing visibility into the actual average trip duration time and how often the bus misses the users desired arrival time.

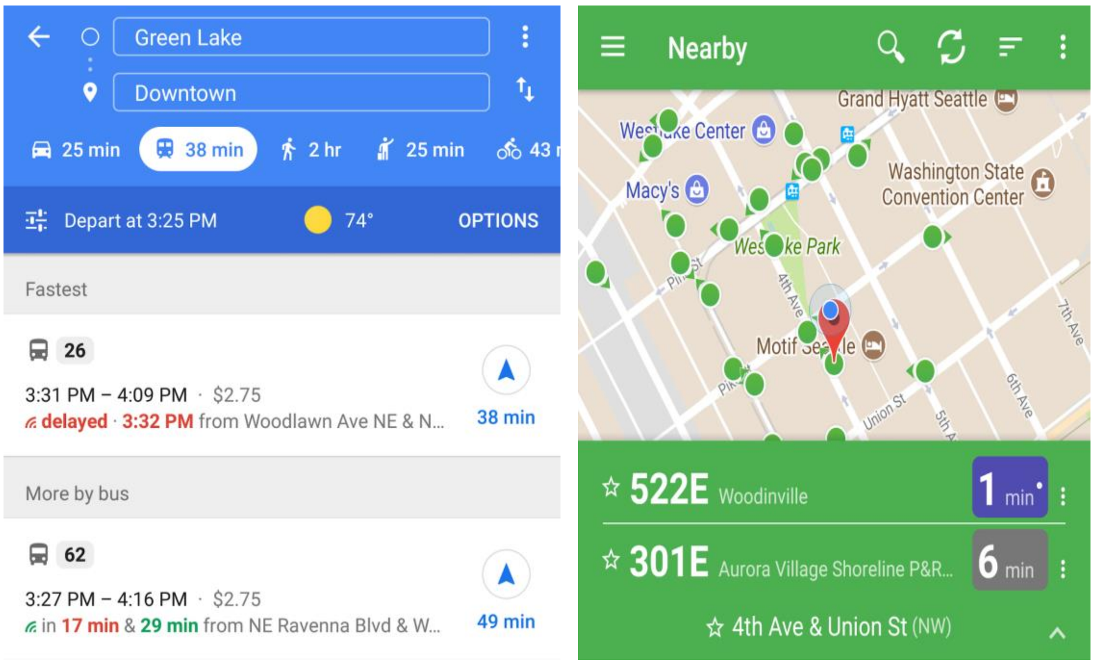


Figure 2 & 3: Google Maps and OneBusAway

Data

The dataset that we used for this visualization was retrieved using the OneBusAway API Service.[3] To leverage the REST API, we were required to request a key from Sound Transit. The REST API does not persist historical data so we scraped a weeks' worth of data, leveraging a python script, between the dates July 28th and August 4th. By limiting our scenario to only the 4 primary bus routes between Green Lake and Downtown Seattle, we reduced the number of bus stops the script needed to retrieve data for to just over 300. After scraping data for a week, the script had produced a soon-to-be-cleaned CSV-file of 574,000 rows totaling 107MB in size.

Each row in the dataset tracked the ID of the stop, route, and trip as well as the epoch milliseconds value for the scheduled arrival and departure times of a route, and the predicted arrival and departure times for a route.

For each stop in our selection of bus routes, we also leveraged the REST API to gather the stop's latitude and longitude. This stop dataset could then be joined together with the data we scraped to provide context on a map as to where each stop was located geographically.

DESIGN DEVELOPMENT PROCESS

First Sketches

We began our design process with a brainstorming session based on the Five Design Sheets (FDS) methodology (Roberts, Headland, & Ritsos, 2015); this process helped us identify what were the core tasks that needed to be supported by our visualization concept. Our initial focus was to track the reliability of a subset of routes at individual bus stops; however, the feedback from brainstorming indicated that it was not an intuitive way to display reliability. After refining our sketches, we concluded that our pilot idea would present a time a user should embark on a given route to reach their destination at the specified arrival time. Providing a map as to where the routes are located and displaying the distribution of how on-time buses would be, were the two recurrent themes in each of our brainstorming pages. The design sheets refined a different element of combining these two ideas. For the first design sheet, we detailed the primary visualization use case of a user specifying a time of arrival and a route. The user would input this information by filtering arrival time through a dropdown menu and clicking on a route plotted on a map. The next design sheet expanded upon this task by having a user select multiple routes through clicking and dragging on the map to evaluate multiple routes and the final sheet focused on evaluation multiple routes within a time window. (Appendix 3)

Methodology

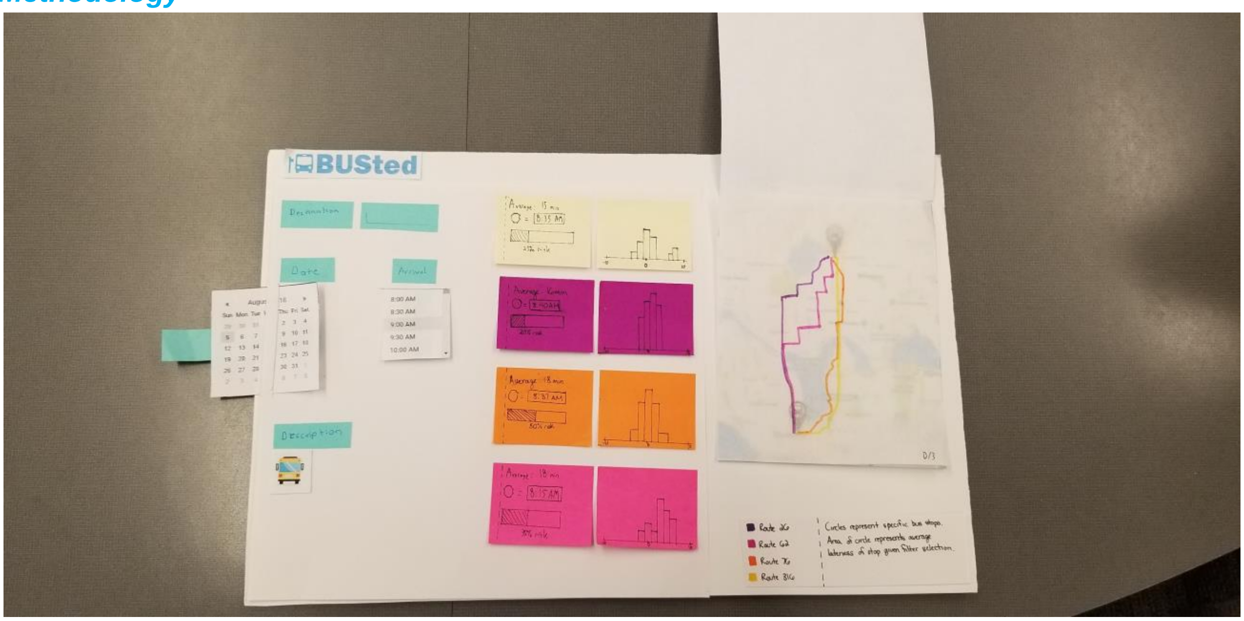


Figure 4: Low-Fidelity Paper Prototype

We developed a low-fidelity paper prototype for usability testing using the Wizard of Oz technique, which involves a team member moving parts of the prototype as if they were the visualization tool responding to user input. For this evaluation, we produced a usability test kit which included three main tasks to present to participants and was complemented by a series of quantitative and qualitative post-interview questions; we requested our participants to Think Out Loud while describing their experience interacting with the prototype.

The paper prototype consisted of four screens, the default screen, and three remaining screens for each of the tasks. The default screen presented on the upper left-hand side the different control filters, and on the bottom, the description for the data (what is the data's value, what are the encodings). A map of the Seattle King County area was located on the right hand-side of the screen and below that a legend displayed the bus routes we had pre-selected for this study identified by their color coding. When a user clicked on the map, we simulated a zoom that would show the different routes. The routes also showcased each stop of the route using circular nodes to identify them on the map. The remaining screens presented the specific illustration for the tasks we wanted to examine; for each task, the screen showed a central section with the different values of the aggregated statistics: the average travel time, the risk factor (of delay time), and a bar chart representing the lateness distribution.

This usability evaluation considered a total of six participants, three pilot tests were done on Wednesday August 8th; the three remaining tests were done with participants that more closely matched our primary persona. Two tests were applied on Saturday August 11th and the third on Monday August 13th, each test lasted a period of 20 to 30 minutes.

Results

Our results helped refine our design. For example, all participants found the low-fidelity prototype visually appealing. Five of our participants mentioned that this concept was engaging, and that they would be interested to use the Tableau version. Four out of six thought our visualization was somewhat intuitive; three concurred the visualization needed to better label the data, five rated the interaction experience as "successful", and all our participants agreed that BUSted enhances the ability to identify bus reliability.

In addition to the usability testing, the post-interview revealed that most of the participants agreed in having a drop-down menu to select the arrival time. All the participants suggested to have clearer labels and one of the participants suggested including seasonal data.

Design Changes

This usability testing process allowed us to gather meaningful user feedback. We concluded what specific points had worked well with our concept, as well as clarified what significant changes were required to begin the development of the Tableau Desktop version:

What worked in the design	Suggested changes
• A horizontal layout with a logical sequence of the stat visualizations for the dashboard. • Setting the filter for each bus route, separate from clicking on the map. • Displaying the map and the option to click on the routes on it. • The color encoding of the risks matching the color of the route. • Using drop down menu to set filters. • The tooltips also showing the data values.	• Change the location of the map to be on left-hand side of the screen. • Locate filters on the top of the screen. Must be seen first to follow an intuitive sequence. • Consider excluding lateness distribution bar-charts. • Clearly specify the risk factor values (0,5,10,15,20 minutes suggested time). • Use appropriate labels to describe the different visualizations shown in the dashboard. • Consider not including data values on the default screen.

There were disadvantages in using paper prototyping. The primary reason being that users did not have a real sense of what clicking on the dashboard in Tableau would be like; we considered this factor could have influenced some of the responses from our participants. Taking this aspect into account, we decided it was important to also gather user perceptions from the Tableau Desktop visualization throughout the development process; and this way, enhance the final design for our visualization.

Cleaning and Exploring Data

Before we could visualize our data we needed to get it into a usable format. A Python script was used to collect the data, and then SQL was used to clean and transform it. The entire scraped dataset consisted of 574,197 rows with 17 variables which was then joined with the stops dataset to obtain the latitudinal and longitudinal coordinates of our stops giving us a total of 19 variables, 10 of which we used for our final visualization.

The data collection method led to excess data that would need to be removed. There were multiple rows of data for the same route, scheduled arrival, and stop with updated "predicted arrival time" values as the route neared the stop. The point where the bus was the closest, "number of stops away", to the stop was used, where its "predicted arrival time" was used as the actual arrival time. Another source of excess data, was the that the 4 routes we tracked extended both north and south of our desired trip, so we filtered stop rows falling outside of our trip.

Since the data is collected by stop we had to piece together completed trips to calculate metrics like average duration. This was done by pairing the vehicle that first started the trip with the vehicle to first reach the end of the route given they had the same "trip id". One anomaly we noticed from this processing was that some of the routes never made it to their destination, due to the route turning into a different route (i.e. the 26 can stay the 26 or turn into the 131 or 132 before reaching our destination stop). These incomplete "trip ids" were removed from our dataset. In the end the dataset was reduced from 574,197 rows to 45,114 rows.

Due to schedule variability, buses are not always scheduled for every stop along a route however the dataset doesn't take this into account and sequences stops based on a particular "trip id". To link our routes in our visualization, the stops needed to re-sequenced to account for some stops being dormant during a particular trip.

Next, a dataset of all combinations of our potential user selected arrival schedules, routes, and directions was created. This was joined this with our completed route data to calculate the average trip duration. We used routes that arrived at the destination within a 2-hour window around the user selected arrival time to build the average. Using the average trip duration, we found the latest scheduled departure time for the user to meet their arrival time. Lastly, the risk percent given the user takes the scheduled departure bus was calculated. We calculated risk as the chance that the user will be 0-5 minutes late, 5-10 minutes late, or 10+ minutes late.

Initial Tableau Implementation (Interactive Prototype)

Our initial implementation in Tableau, shown below, closely mirrors our paper prototype. The initial view presents all routes, for all days of the week at 1:00 PM. While this view only provides information for a limited use case, three filters are provided for interaction to filter down the time and day of the week a user is interested in taking the bus, as well as, which direction they are headed in. When a user selects a given destination arrival time, data that is both from an hour later and an hour earlier is also filtered in due to our limited data set. This visualization also supports multiple coordinated views. Selecting a route from the map or the statistics will highlight the selected route.

This implementation differs from the designs generated from the Five Design Sheets exercise we performed in class, primarily due to layout restrictions in Tableau. Instead of stacking each statistic vertically per route, they are laid out horizontally as each statistic is captured in its own sheet that is scaled to fake each collection of data points being presented on what looks like a single row.

Figure 5: Initial Tableau Desktop Implementation

User Testing on Tableau Desktop

In addition to the pilot testing that was performed in class on the low-fidelity paper prototype, we also engaged three users who matched the profile of our persona to test on a prototype built in Tableau. Each of the three users were presented with the refined tasks that had been updated from feedback given during the pilot tests. These tests were performed individually by members of our team.

The results of our second round of testing were more positive than the first. Given the staggered nature of these user studies, we were able to refine elements of the visualizations that users found confusing between tests.

The portion of the visualization that users have expressed the most confusion about, was how we presented the calculation for bus reliability. After the pilot test, this calculation changed from being a distribution of bus lateness to being a percent chance that the bus would be N-minutes late or more where the user was able to select N from a list of integers. Users still found this to be confusing, so our next idea was to provide cumulative intervals for risk so users could evaluate if the bus was zero or more minutes late, or five or more minutes late. In both of the last two user tests we performed, the users interpreted these cumulative ranges as finite ranges. For example, they thought that the value 0 was indicative of the range 0-5. Given this feedback our final design embodies the users' intuition.

Figure 6: Progression of Risk Design/Display

Evaluation of the Final Visualization

Within this section we will be evaluating our final visualization, how well it lives up to the concept proposal, how the concept changed due to user feedback, and how well we have applied concepts from the course. Our final visualization can be viewed at the following URL:

https://public.tableau.com/profile/sean.miller2492#!/vizhome/BUSted/BUSted_1

Project Goal and User Tasks

The goal of our project was to empower users to understand the reliability of different bus routes and to aid them in making better decisions around which bus they can take. While users were able to complete each task within the Tableau dashboard, feedback indicated that further flexibility of filtering would be appreciated to better explore and comprehend the visualized data.

Information Visualization

Within this section we will take both an all-encompassing view of the information visualization we produced as well as dissecting each part that comprises the whole view.

The expected user flow highlights different tasks from both the three categories of tasks presented in Heer and Schneiderman's article on interactive dynamics for visual analysis.[4] The primary means of interaction focuses on the filter task. Users select the direction they are headed in, the day of their trip, and what time they need to arrive and the visualization filters the view based on their selections. After the view has been filtered, users can then select items in either the map or in the aggregated stats and the coordinated views will reflect their selection by filtering down to the selected items. The Tableau Public interface records a user's actions and allows them to undo or redo as well as to reset the visualization to its default state.

Route Map

The Route Map provides the user with an overview of the Greater Seattle Area while outlining the path each route takes. Each route is assigned a separate color to allow pre-attentive processing to handle segmenting each of the paths into distinct entities. The map also supports the idea of multiple coordinated views as selections on the map through brushing or clicking are reflected in other views on the dashboard.

Figure 7: Route Map

Upon initial presentation the map is filtered down to 9:00 AM on August 15th. Due to limitations in Tableau's design, parameters cannot be made dynamic so we chose the default to be the day of this class' final presentation. Points are plotted on the map based on a stop's latitude and longitude and then are connected in the order the bus route takes. While this is not a perfect mapping for the streets or highways a bus can take it is a good enough approximation. Even given our focused use case, the route for the 316 bus was occluded by the route for the 76 bus and needed to be shifted to account for this. Given a larger data set, the initial view presented would be much more cluttered and different approaches would need to be taken to address scaling the number of routes considered.

The map leverages the tooltip function of Tableau desktop to provide details on demand. When hovering over a given route on the map the tooltip displays the route number, what information is presented on the sign of the bus and provides a link to the King County Metro site for that route.

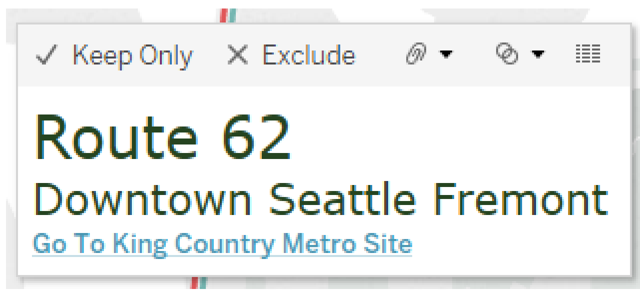


Figure 8: Tooltip From Hovering Over Route on Route Map

Aggregate Statistics

The Aggregate Statistics view consists of two text fields and a bar chart. The chart highlights the chance that for a given set of user-set filters that a bus is more than the displayed number of minutes late.

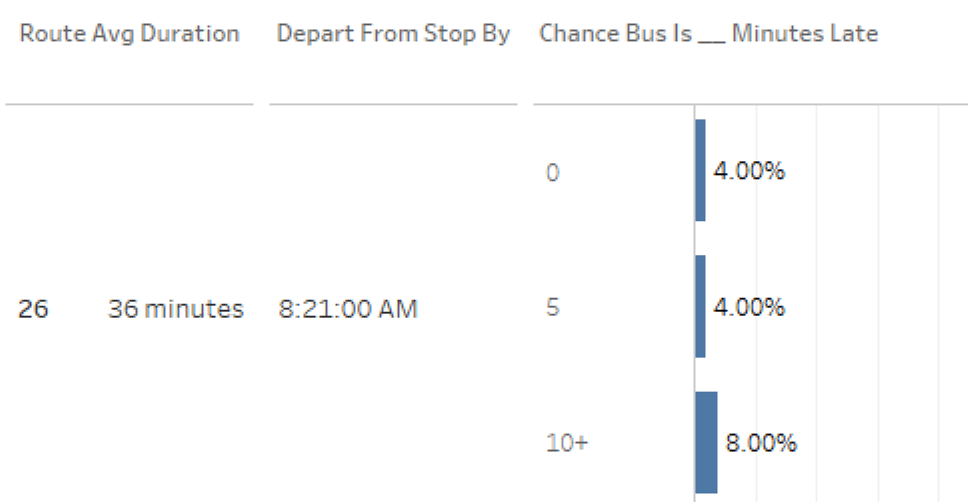


Figure 9: Route 26 Aggregate Statistics

For each route, the first value presented is the average duration of the bus route for the selected arrival time from our data set. This duration can deviate from the scheduled time a trip takes and the suggested “Depart From Stop By**”** time is adjusted to reflect the bus being faster or slower than expected. Originally, a bar chart at the end of a row presented a series of ratios that expressed how likely a bus route was to be 0, 5, or 10 minutes late or more. During the user testing, two users found this metric to be confusing as their expectation for each row in the bar chart was for the value to be indicative of a range. For example, the 0 minutes value indicated the range 0 to 4.9 minutes and the 5 minutes value indicated the range 5 to 9.9 minutes. The current iteration of the chart reflects this feedback.

The colors of the bar chart reflect the color of the route on the Route Map. This enables easier association between routes on the map and the reliability of that route. Each row is prepended by the route number to give context to the statistics if presented without the Route Map. As Edward Tufte suggests in *The Visual Display of Quantitative Information* that “the representation of numbers… should be directly proportional to the numerical quantities represented,” we have scaled the axes of the bar chart to always be 0-100% to be consistent and comparable.[5]

Filters

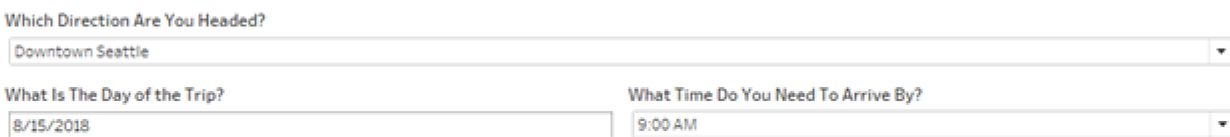


Figure 10: Filters for BUSted

The filters for direction of the trip and the time the user needs to arrive by are both dropdown menus. While the direction menu supports the use case we present within our visualization, adding every possible combination of destination start and end points would provide undue cognitive load to users. Systems such as Google Maps can use addresses or device location to map the route nearest to the user. Unfortunately, such features are not provided within Tableau.

When selecting the day of the trip, users are presented with a calendar. Feedback from one user test recommended that the visualization allow them to select several days which is not provided functionality for date parameters in Tableau. This could be worked around by providing multiple parameters for a start and end date but we decided that was outside the scope for our user tasks.

Future Work

Our current information visualization was constrained to 4 King County routes and limited by a weeks’ worth of data from OneBusAway. Future work and resources would focus on increasing map coverage with a larger set of historical data, as well as collecting ridership numbers. With this added complexity, we would be able to look into transit seasonality patterns, which was part of the feedback we had received on our low-fidelity prototype. Additionally, we could look into more systemic issues like poor bus route coverage or frequency to help increase overall reliability for bus commuters.

In addition to the constraints provided by the sample of data we scraped, we were also limited by the interface that Tableau Desktop provides. When users interact with Google Maps to find bus routes to take, they input an address and the application is able to map that address to nearby bus stops. Tableau does not provide this feature. In the future, we would look into alternative methods of presenting our visualization which allow a greater range of user interaction. These methods would likely be a phone application or on a webpage leveraging a technology like D3.js.

REFERENCES

1. https://www.seattletimes.com/seattle-news/transportation/your-metro-bus-is-on-time-more-often-than-you-may-realize-heres-why/
2. http://bigbytes.mobyus.com/commute.aspx?County=13&State=8&MinMiles=20&MaxMiles=100&MaxCoV=0.8&CountyType=work
3. http://pugetsound.onebusaway.org/p/OneBusAwayApiService.action
4. Heer, J & Shneiderman, B & Park, C. (2012). A taxonomy of tools that support the fluent and flexible use of visualizations. Interactive Dynamic Visual Analysis 10. 1-26.
5. E.R. Tufte, The Visual Display of Quantitative Information, Cheshire, CT: Graphics Press, 1983.

Appendix 1 – Primary Persona

OLIVIA Recent UW Grad

"I like to always arrive on-time"

Age: 22 Occupation: Customer Marketing Location: Seattle, WA

ABOUT OLIVIA

+Olivia is a recent UW grad and she just landed an entry-level job position in a tech company in downtown Seattle.

+She moved in with her brother who lives in Green Lake to save money for a year.

+She doesn't own a car, so for now all her commuting is by bus.

CASE SCENARIO

+She loves that her new job will be mostly remote; however, she will still have to work in the office several days a week.

+She values punctuality but having to commute by bus in Seattle is always a challenge, especially if during rush hour.

+She usually uses one bus away app to track her bus arrival time, but she has experienced that the information provided sometimes is inaccurate. There have been days where she has waited more than 30 min for a bus.

+She really hates having to take extra time for her commute just to make it on time.

+She would like to know what bus routes could be more reliable during the time she has to commute.

+Doing some research online, she comes across BUSted, an online visualization tool that will enable her to know the historical reliability of a bus and help plan her commute more accurately.

MOTIVATIONS

+Not having to take extra time for my commute.
+Arrive on time to my destination.
+Know which bus is the fastest route.

PRIORITIES

DEAL BREAKER

+Inaccuracy of the tool

WOW FACTOR

+Arriving on time to my destination will allow me to make a better use of that extra time I always take for my commute. It would be great to use that time for other activities during the day.

APPENDIX 2 – DATA

Example of the Raw Data Format:

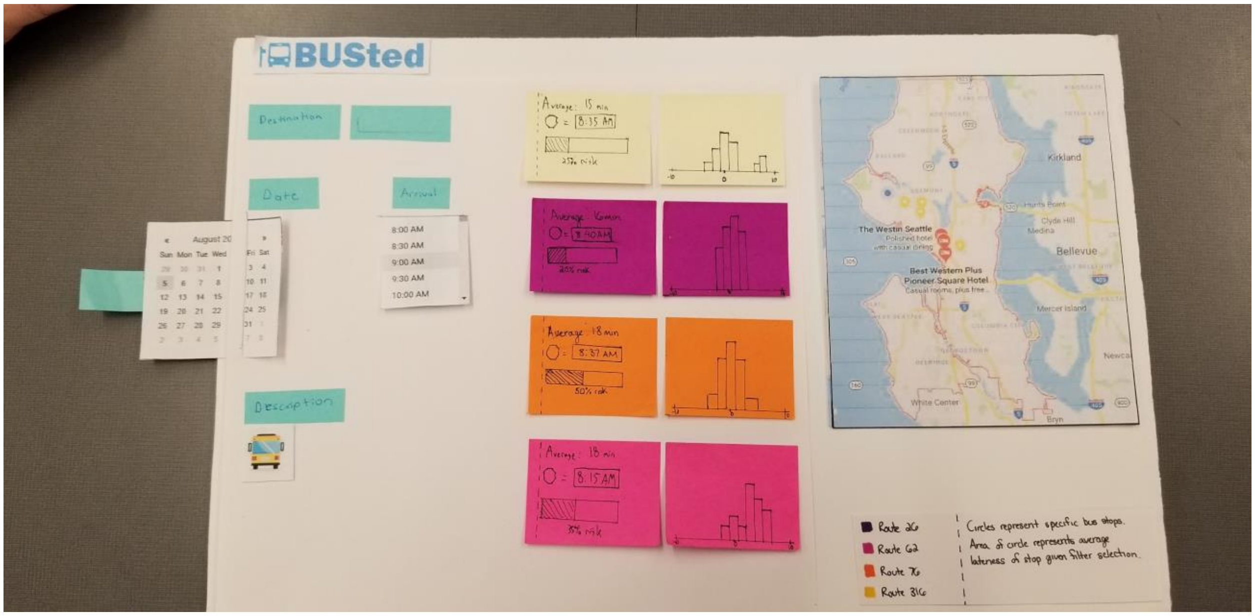
Field	Value
stop_id	1_17022
trip_id	1_39433420
route_id	1_100151
stop_sequence	22
scheduled_arrival_time	1532815081000
predicted_arrival_time	1532815830000
number_of_stops_away	7
trip_headsign	Northgate Transit Center East Green Lake

Example of the Stop Data Format:

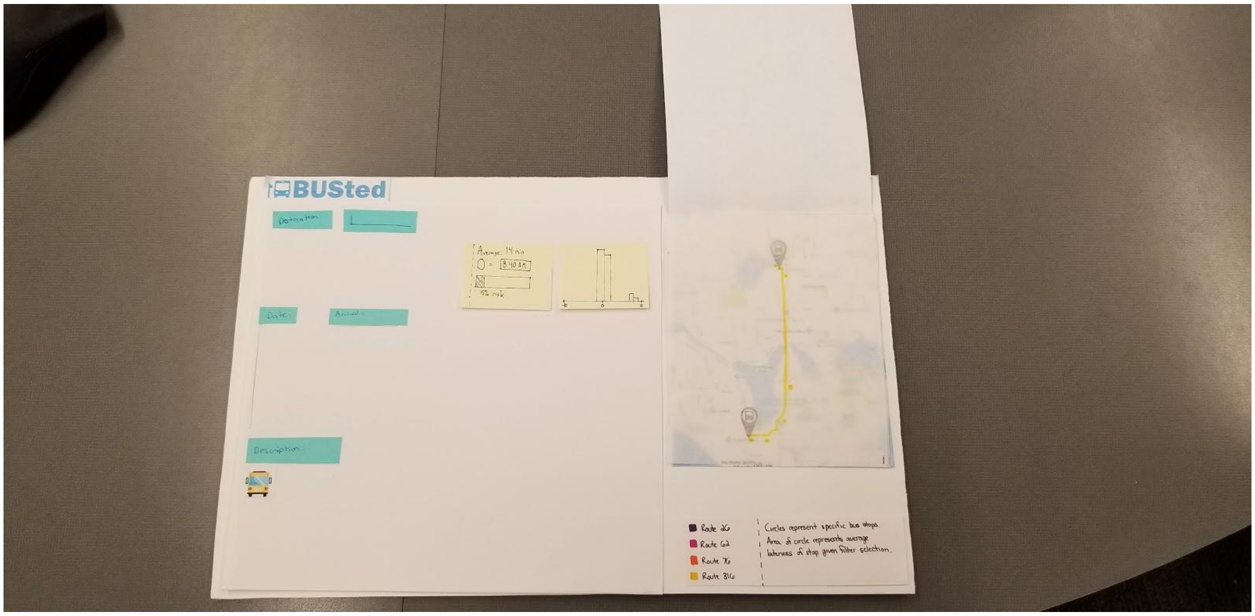
Field	Value
stop_id	1_17022
route_id	1_100151
lat	47.695854
lon	-122.330673
direction	E

Appendix 3 – Low-Fidelity Prototype

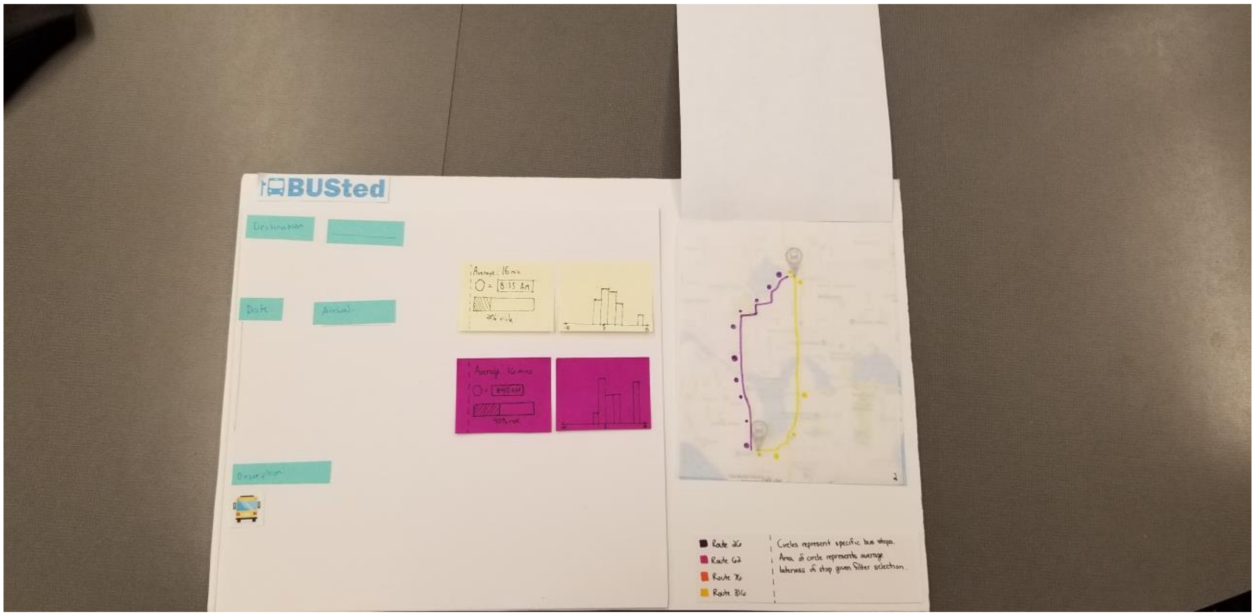
Default Screen



Task 1



Task 2



Task 3