# Pasadena Transit: Leveraging big data insights to improve public transportation

James Blackwood[1], Coauthor 1[1]*, Coauthor 2[1]

**Abstract**
Metropolitan transportation systems are a field well-suited for the application of robust data analysis. Determining who uses public transportation, how a system generates revenue, and how a system can be improved are central questions for municipalities across the world. By leveraging data insights, cities can improve their transportation systems and improve the well being of residents. Pasadena Transit (formerly Pasadena Area Rapid Transit, or Pasadena ARTS) is the city of Pasadena's public transport system, consisting of eight bus routes servicing the whole city. In this paper, we utilize machine learning techniques to analyze data from the Pasadena Transit system. We then use these results to inform policy decisions. In particular, we recommend removal of the 51/52 routes, a targeted advertising campaign to the affluent population of Pasadena, and a possible partnership with Uber/Lyft to support Sierra Madre Villa (SMV) Station transit.

[1] *California Institute of Technology*
*__Corresponding author__: clee@caltech.edu
Our web app is here: https://pasadena-area-transport-system.herokuapp.com

## Contents



## 1. Problem Statement

We begin by providing context and benchmarks for Pasadena Transit operations. Every fiscal year, the City of Pasadena allocates a limited amount of funds towards marketing and improving Pasadena Transit (formerly Pasadena ARTS). The proposed 2017 transit budget sits at $9,606,000, increasing from the current year's budget of $9,176,000 by 5.4 %. However, this difference is due to a projected increase in hours of operations and hourly rates across the Department of Transportation. As such, a prudent assumption is that the Transit Service's operating budget in 2017 will be the same as the current year's and is an important consideration when making policy recommendations.

An ongoing challenge regarding the Pasadena Transit services is to increase ridership. For example, in 2015, efforts to re-brand the transit system and optimize bus schedules for the Metro Gold Line corresponded with a 5% increase in ridership. In order to improve on this trend, insights from data are necessary. In particular, we consider three key lines of analysis:

- *Ridership*: Which type of boardings take place. Also, where do they occur and what are the demographics?
- *Routing and Timing*: What are the most popular routes and popular stops? Are route delays and ridership a function of the time?
- *Profitability and Impact*: What is the net profitability

of each route and the system at-large? How can the system better serve the city?

## 2. Data Set

The data ensemble used in analysis was presented in two distinct forms: TAP data and GPS data. These data sets respectively contained information about TAP user information and bus routes of the Pasadena transit system over approximately three weeks (4/11/16 - 4/30/16).

### 2.1 Raw Data

#### 2.1.1 TAP Datasets

A single TAP transaction was represented as a row with 16 attributes in the TAP CSV files. The key attributes were Date, Fare Product, Transit Card ID, and Bus Number, where Date represents the precise time of transaction, Fare Product represents the specific type of transaction, Transit Card corresponds to a unique user ID, and Bus number denotes a specific bus ID on which the transaction took place. We refer to all data of this form as "TAP data."

#### 2.1.2 GPS Datasets

Each row in the GPS datasets represents a trip taken by bus. These trips are uniquely identified by its 11 attributes, 7 of which (Vehicle, Hour, Minute, Route, Departure, Arrival, Trip No) are of importance to this study. We refer to all data of this form as "GPS data".

#### 2.1.3 Bus Timetables

The bus timetables contained information pertaining to the route schedules. The schedules for each route varied depending on whether it was the weekday or weekend and displayed the arrival time at each station along the route.

#### 2.1.4 Income Data

An estimate for the median income in the surrounding area of each bus stop was gathered from Census Explorer, an interactive platform created by the U.S. Census Bureau. Many of the stops were located at the intersection of several census tracts. As a result, the median income was estimated by weighting the median incomes of the neighboring census tracts by their respective populations. The income data was then combined with the merged data set according to the individual stops.

### 2.2 Data Pre-processing

#### 2.2.1 TAP/GPS Aggregation

We note that the rows of the TAP and GPS datasets share common attributes. Therefore, we performed a relational join on the two datasets where corresponding attributes were equal to one another. Note that only Vehicle - BusNumber, and Date - Date are perfectly equivalent. The time of the Date attribute in the TAP data must fall within the interval spanned by Departure and Arrival. Because of this, the relational join did not result in a perfect mapping of TAP to GPS data. There exist a few cases where a TAP transaction was mapped to more than one bus trip. The frequency of such correspondences was rare, and essentially serves as noise. Results indicate that this noise was insignificant.

The aggregated data consisted of rows representing TAP transactions with a total of 27 attributes from both TAP and GPS data.

#### 2.2.2 Station Prediction Algorithm

Using bus timetables, and the aggregated data, the station that a user departed from could be predicted with high accuracy. To determine this, we minimized the difference between a bus's departure time and the time a user tapped on. We predicted the station based on the minimum difference. With this data, we now had the ability to determine the type of transactions that occur at each station.

## 3. Analysis and Discussion

### 3.1 Routing and Timing

#### 3.1.1 Linear Model of Delay

To ascertain if delay times are a function of time and route, we modeled departure delay time of a transaction as a linear function of time of day and route. The results of the OLS regression model are presented in Table 1. We recovered 3 statistically significant regression parameters (at the 5% level) and indeed the F-test implies strong rejection of the null hypothesis that there is dependence on time and route are jointly zero. We note that intercept, and the dummy encoders for routes 10 and 20cc are individually significant. Note that the intercept term accounts for a sizable portion of the delay in magnitude, however, route 10 is statistically late, and route 20cc is statistically earlier

From this model we recovered two salient results. First, we can posit—the somewhat counter intuitive result—that time of day does not affect delays. The second result was that route 10 runs statistically late.

#### 3.1.2 Bus Stops

We then analyzed specific stops in the system. We looked to classify the bus stops as—nominally— "valuable" and "not valuable." To do so, we defined "value" to be a function of the average delay for all transcations and the ridership count. We clustered the stops in these two dimensions. To correct for skewness, we used the natural log of the passengers. We decided to use a spectral clustering approach because the method is well suited for situations calling for a low number of clusters. Since we were looking to impose a binary cluster set, we thought spectral clustering was the best option.
The results from the clustering were both clear and intuitive. We note that passenger load dominates the clustering, yet it did not uniquely determine the clusters. We present this in Figure 1. We provide labels to give an indication of outlier stations that we classify as "not-valuable."

### 3.2 Ridership Breakdown

Analysis of ridership breakdown was done under the assumption that fare product is a distinguishing factor among users.

**Table 1.** Regression output from linear model of delay.
Note that here a positive coefficient corresponds to an *increase* in delay.

| Variable | Coeff. | Std. Err. | t-stat | p-value |
|---|---|---|---|---|
| Time | 0.1330 | 0.2718 | 0.49 | 0.6246 |
| 20cc | -1.9015 | 0.6425 | -2.96 | 0.0031 |
| 51 | 0.5690 | 0.6562 | 0.87 | 0.3859 |
| 52 | -0.2182 | 0.8220 | -0.27 | 0.7906 |
| 60 | 0.0105 | 0.6981 | 0.02 | 0.9880 |
| 32 | -0.1004 | 0.6506 | -0.15 | 0.8774 |
| 31 | -0.7666 | 0.6453 | -1.19 | 0.2349 |
| 20cw | 0.4325 | 0.6422 | 0.67 | 0.5006 |
| 10 | 1.2803 | 0.6450 | 1.99 | 0.0472 |
| 40 | 0.3939 | 0.6441 | 0.61 | 0.5408 |
| Intercept | 2.7794 | 0.6427 | 4.32 | 0.0000 |

$R^2$: 0.1083; Adj. $R^2$: 0.1077
F(10, 13790):167.5279; p-value: 0.0000

An online heat map was developed to allow easy visualization of the activities of different fare product holders at various times of day (Figure 6).

#### 3.2.1 Fare Product: Holistic View

While the dataset does not explicitly state the demographics and background information about the transit riders, it implicitly denotes age groups and people with disabilities through the type of fare product the riders individually use. In order to increase revenue and ridership, it is necessary to know who the main customers are, and who needs more convincing to use the transit systems.

We start off with a holistic view of the distribution of fare product used from 4/11/16 - 4/30/16. From Figure 2, it is clear that the vast majority of transit users who pay by TAP (rather than cash) are are ASI (Access Service ID) holders. Regular TAP (Reg SV Regular) users come in second, followed by senior citizens or those with disabilities (Reg SV Sr/Dis).

#### 3.2.2 Fare Product: Station Breakdown

Having seen which fare products are most popular among transit customers, we then observe the top three most prevalent fare products in context of the different stations to determine if certain stations are used more heavily by customers holding certain fare products. From Figure 3, it is clear that Raymond and Holly is the highest frequented station among ASI, Reg SV Regular, and Reg SV Sr/Dis holders. The second most frequented station among ASI users is Fair Oaks and Washington; however, this station does not see much use by the other two pass holders. The majority of Reg SV Regular and Reg SV Sr/Dis holders depart from Lake and Colorado, and Arroyo Pkwy and California.

This analysis shows clustering among three different fare product users at various stations. One possible explanation for this behavior is that users holding certain fare products tend to live in the same general area. For example, students may tend towards areas with cheaper rental costs.

#### 3.2.3 Fare Product: Route Breakdown

Next, we look at fare product breakdown by route. The general trend seems to show that routes 10, 20cc, and 20cw are used most among all ASI, Reg SV Regular, and Reg SV Sr/Dis holders (Figure 4).

#### 3.2.4 Ridership: The effect of the surrounding area

To determine the effect of an area on ridership, we posited that the income of a surrounding area can uniquely determine the ridership of the nearby bus stops. The functional form of this model is therefore given by

$$R_{stop} = \alpha + \beta\ log(I_{med}), \tag{1}$$

where $R_{stop}$ is the total revenue generated at a stop, and $I_{med}$ is the median income of the census region in which the stop is located. We present the results of this model in Table 2.

We recover a statistically significant regression coefficient, as well as a robust economic effect. For a 50% increase in the surrounding median income, we recover a $4000 annual decrease in TAP revenue. Given the overall small degree of TAP ridership relative to overall ridership, this result is actually quite significant and suggests that wealthier areas ride the bus with much less frequency. While this implication is intuitive, the degree to which it appears to hold in Pasadena is striking.

There also appeared to be potential heteroskedasticty in the data. In particular, the *variability* in ridership levels seemed to decrease as the income off the area increased, in addition to the decrease in magnitude. We therefore performed both a White's Test and a Breusch-Pagan Test to test for heteroskedasticity. The F-test significance of the tests were 0.11 and 0.12 respectively, and we thus verified the assumption that our data was homoskedastic and we accepted the results of our OLS regression.

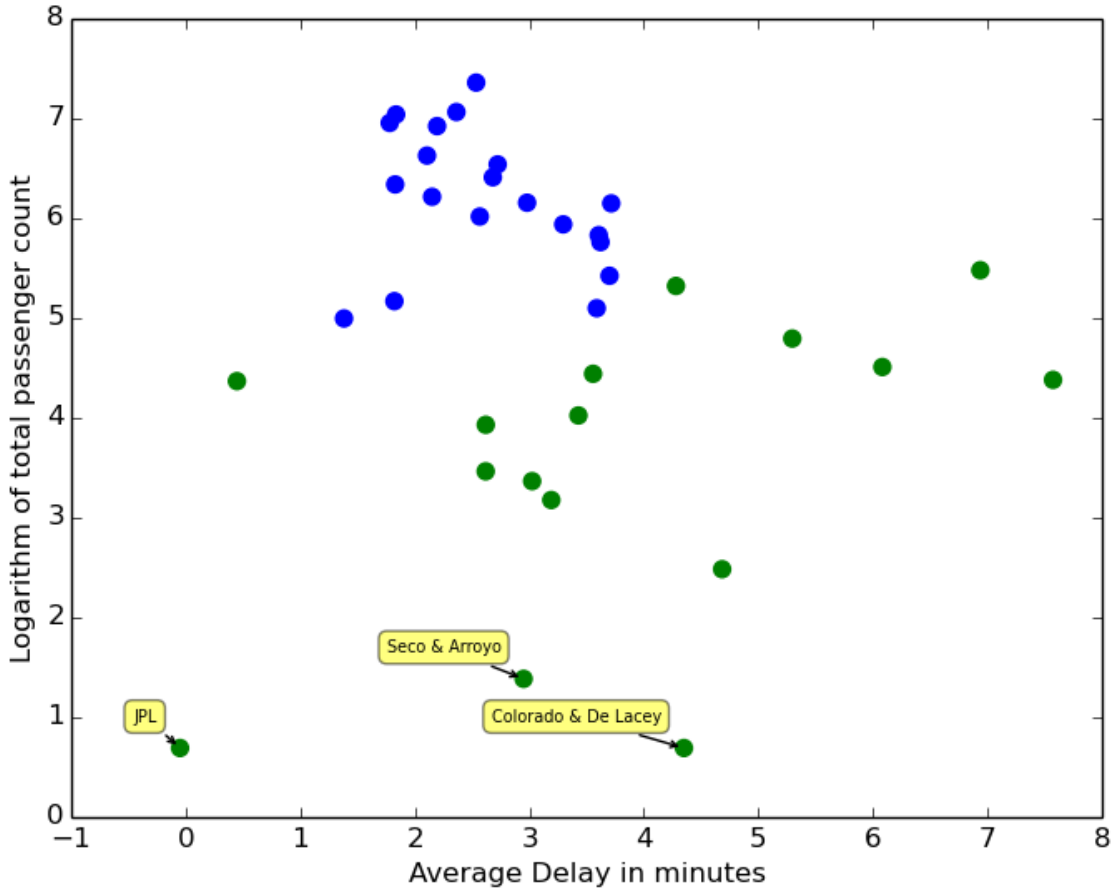


**Figure 1.** Spectral clustering of bus stops on passenger load and delay. The Radial Bias kernel was used to determine affinity matrices. Clustering results were intuitive and provided insight into under-loaded and perennially late bus, or conversely, stops that are carry a high passenger load but do not choke the system.

## 4. Conclusions and Recommendations

### 4.1 Partnering with Uber/Lyft

From Figure 6, it is quite clear that although SMV Station shows very significant traffic, the surrounding stops show very little traffic. We propose that Pasadena partners with Uber/Lyft in order to provide a discount to riders traveling to or from SMV station. The City of Pasadena would help to subsidize this discount. Although more analysis is necessary, we believe that this might be able to increase profitability due to the decrease in gas costs and the ability to completely remove the low-performing bus stops from Pasadena Transit routes entirely. This model would allow riders from SMV station to still find transit routes, but would remove some of the inefficiencies currently in Pasadena Transit routes. The precedent of such a model comes from the City of Altamonte Springs in Florida which received additional outside assistance to cover the subsidy costs.

### 4.2 Improving On-Time Performance

One of the performance measures cited by the City of Pasadena for the Transportation Program is the on-time performance of the Pasadena Transit. Although the FY 2016 mid-year performance of 94% actually exceeds the FY 2016 goal of 88%, we believe that our analysis can greatly improve the on-time performance of Pasadena Transit. We believe that completely excising the 51/52 route would greatly improve both the on-time performance as well as the overall profitability of the Pasadena Transit. All but one of the stops on the 51/52 are categorized into the poor-performance cluster according to Figure 1. These are the stops which show either very low passenger count, very high average delay, or a combination of the two. As a result, eliminating these stops improve an even more important metric of on-time performance which accounts for the variability in passenger load of different bus stops. Additionally, these routes could be run only during special events (e.g., Rose Bowl Game), if they are believed to be critical during these times. Although completely eliminat-

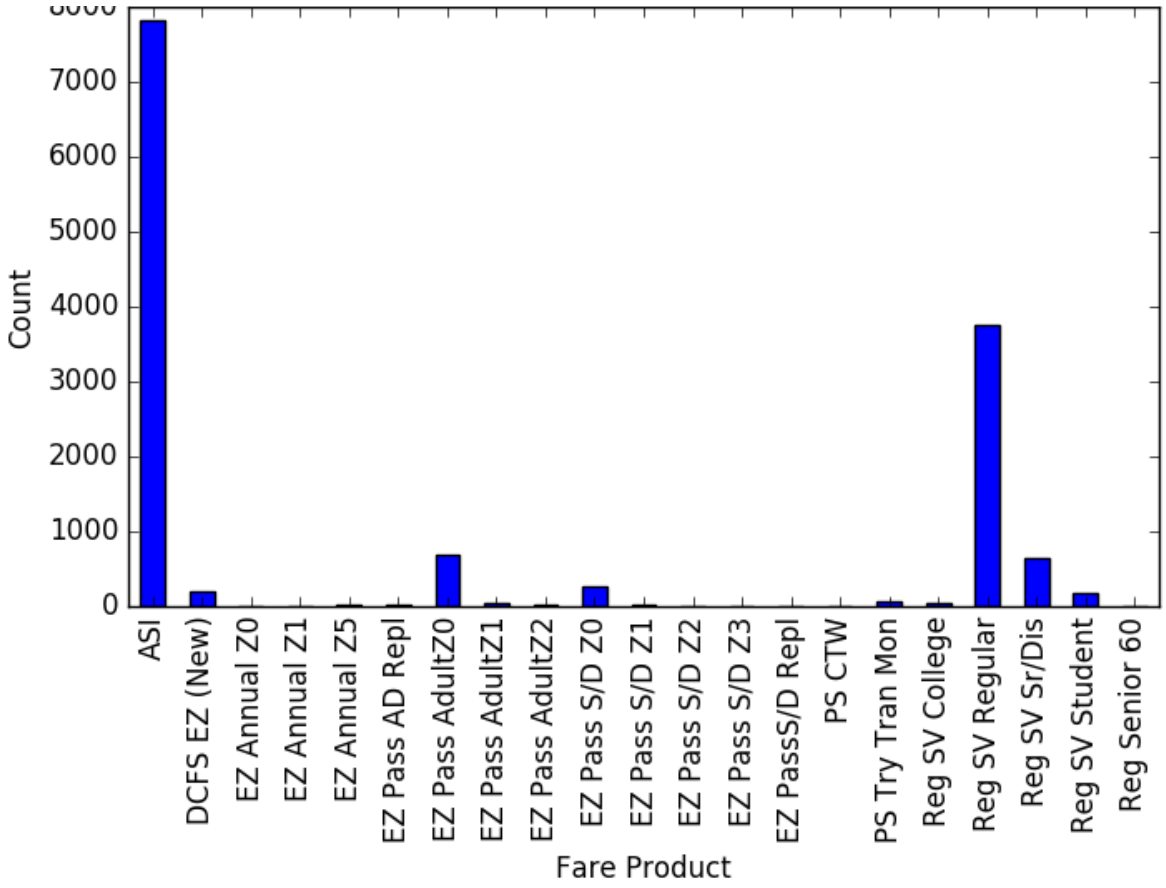


**Figure 2.** Histogram of different type of fare products used from 4/11/16 - 4/30/16. Clearly, holders of ASI (Access Service ID) are more prevalent than the other fare product holders. After ASI, regular TAP card holders are the next most prevalent with a count of around 3500.

ing these routes might be quite detrimental to the riders, we believe that these routes are primarily used by JPL employees or ArtCenter College of Design students. If these institutions view the transportation routes to be of enormous significance, they can help subsidize them or arrange alternative transportation methods (e.g., shuttles).

### 4.3 Improving Ridership in Affluent Areas

One of the statistically significant trends we identified is that bus stops in surrounding areas with higher median incomes tend to have significantly less ridership (and therefore TAP revenue) than bus stops in surrounding areas with lower median incomes. This is an important result because it identifies a segment of the population that the City of Pasadena can identify as a segment with potential growth in ridership. Although the reasons for lower ridership by affluent individuals should be confirmed through market research, we assume that this is simply a combination of stigma and transportation alternatives. In order to address this problem, we recommend an advertising campaign targeted at the affluent population of Pasadena. This campaign could focus on the impact an individual could have by switching to public transportation to visit Old Pasadena on the weekends instead of driving. This impact could be quantified by some salient metric such as reduction of fossil fuel emissions.

## 5. Appendix A: Exploratory Results

We present briefly our exploratory results, in which we tried to apply more novel data analysis techniques. These results produced non-salient results or results that do not have an intuitive interpretation. Therefore, we present them as exploratory, rather than primary analyses.

### 5.1 K-Means Clustering

In more exploratory analysis, we performed clustering analysis on fare product. In Figure 5, we present our clusters transformed into the first two principal components. It is important to note that these first two principal components only explain 8% of the variance in the data, so the dimensional reduction is not too insightful in terms of the visualization below. However, we recovered some robust clusters, and thus this model can help classify customers in the future in terms of how they pay.

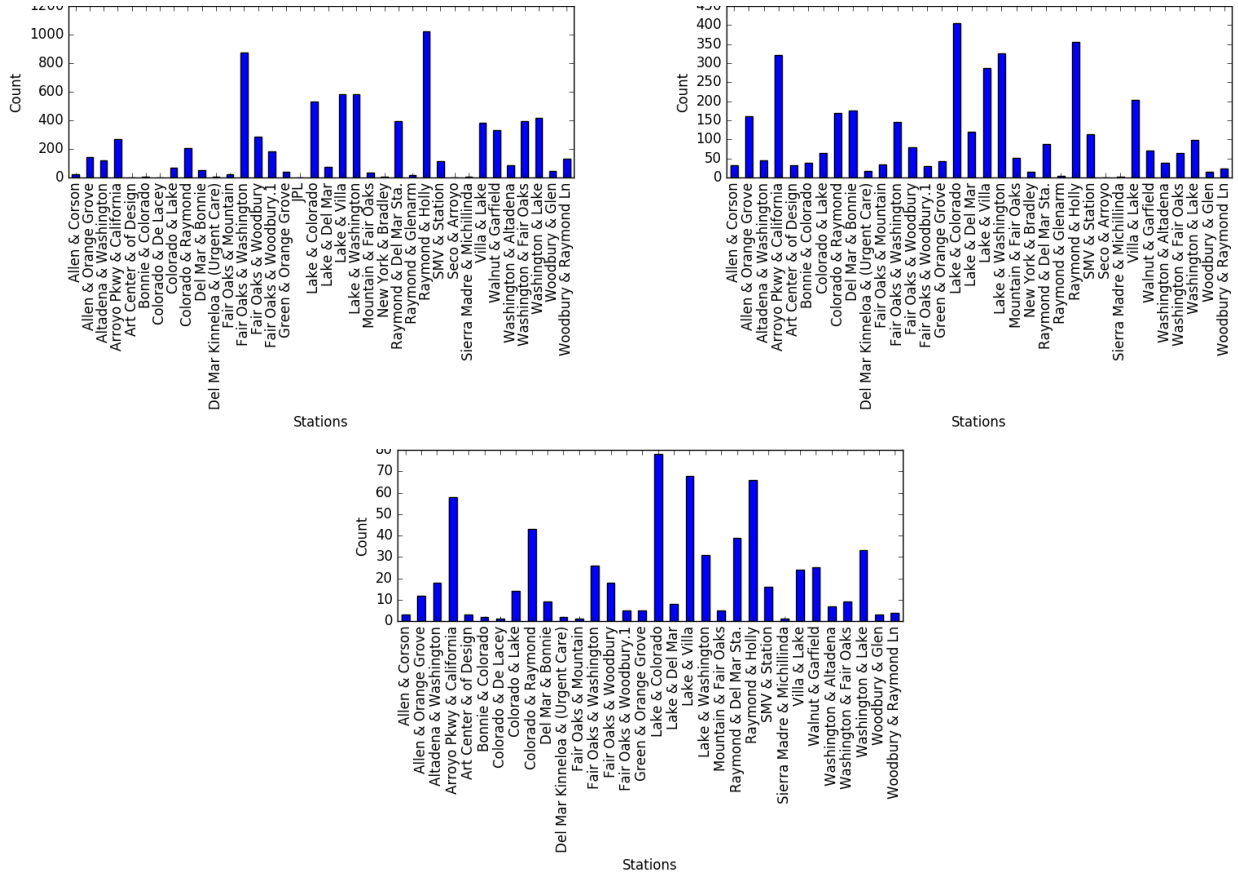


**Figure 3.** Fare product breakdown by stations. The graphs above display the breakdown of ASI, Reg SV Regular, and Reg SV Sr/Dis used at different stations. By inspection, different fare product holders seem to be biased towards departing from certain stations. Stations not shown had no users of the fare product of interest.

### 5.2 Contextual Bandits

The contextual bandit model was applied in attempts to determine time-dependent riding habits of users in addition to building a robust route prediction tool for the City of Pasadena. However, because the average regret was around 0.8, the model does not accurately predict ridership behavior. This may be due to the fact that only about 3 weeks of data was used by this algorithm. Because rider behavior may recur weekly rather than daily, not enough samples may have been seen for each user.

## Acknowledgments

We extend our gratitude to Sebastian Hernandez for his support throughout this project. We would also like to thank Alice Lin and Colin Camerer.

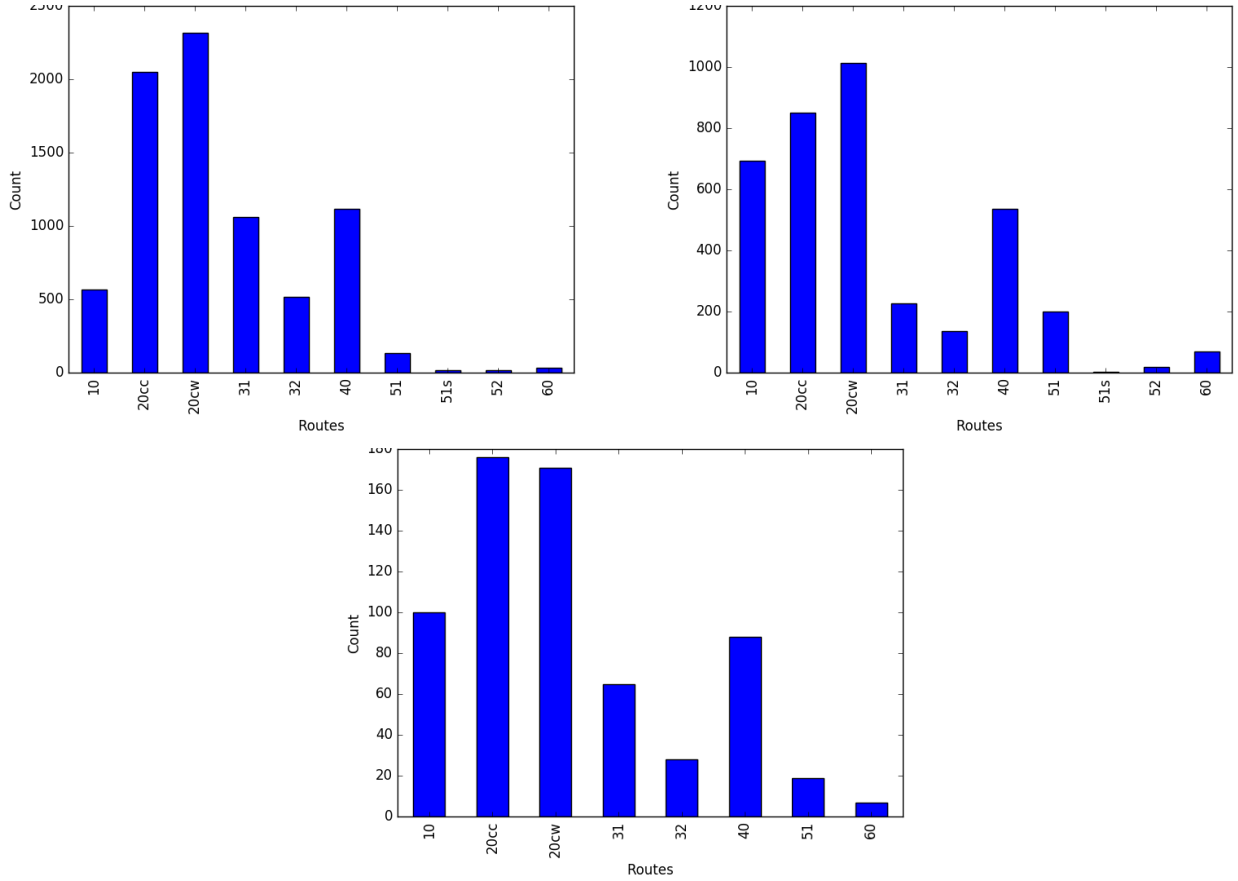


**Figure 4.** Fare product breakdown by route. The fare products used on the routes generally seem to have their maximums reached at common routes (10, 20cc, 20cw). Note that routes that aren't shown on the plot had no users of the fare product of interest.

**Table 2.** Regression output from linear model of TAP revenue

| Variable | Coeff. | Std. Err. | t-stat | p-value |
|---|---|---|---|---|
| Log Median Income | -536.65 | 189.37 | -2.83 | 0.0088 |
| Intercept | 2752.57 | 923.40 | 2.98 | 0.0062 |

$R^2$: 0.2360; Adj. $R^2$: 0.2067

F(1, 26):8.031; p-value: 0.0088

**Figure 5.** K-means clustering of Fare Product transformed on the first 2 principal components. There is no intuitive interpretation for the first two principal components and indeed they only explain 8% of the overall variance in the data. We still recover some dominant clusters and this model may help in characterizing future customers.

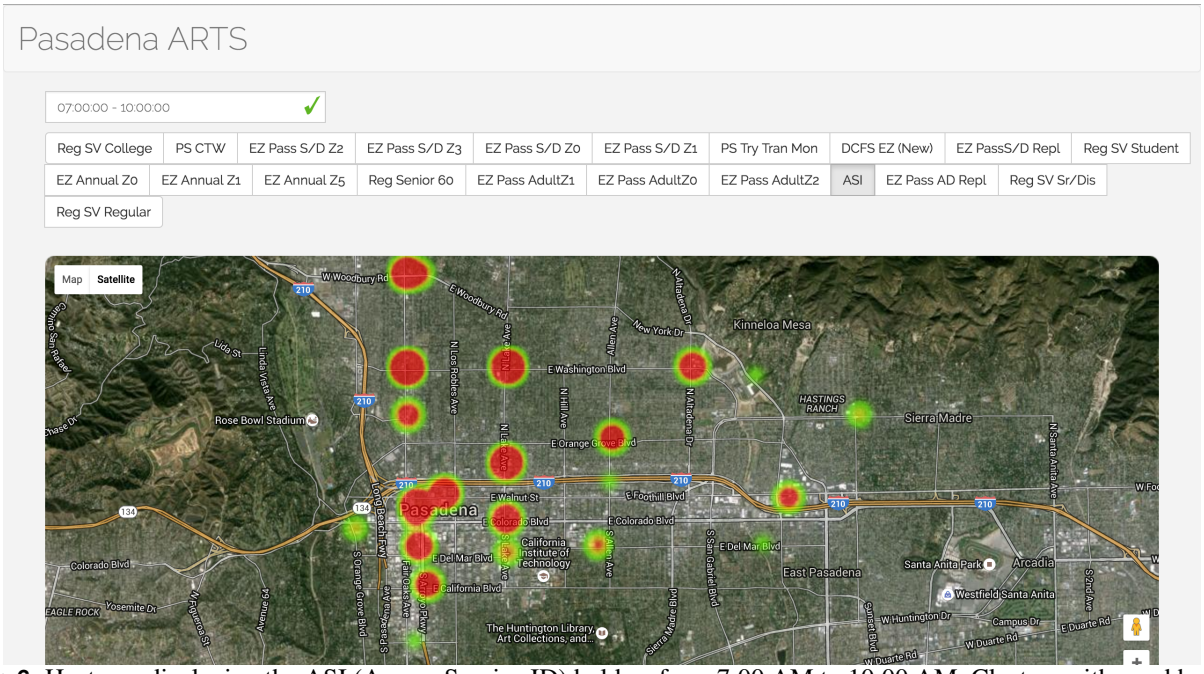


**Figure 6.** Heatmap displaying the ASI (Access Service ID) holders from 7:00 AM to 10:00 AM. Clusters with a red hue denotes higher concentration of users at the station.