Dennis Silva, Hang Yu, Neel Gehlot, Qiuyi Hong, Yinkai Ma, Qian Xu

Abstract

The aim of this report is to utilize business and data mining strategies to provide tailored recommendations to SFO that will enhance customer experience and create the foundation for future business strategies.

customer Suvery data analysis

San Francisco International Airport (SFO)

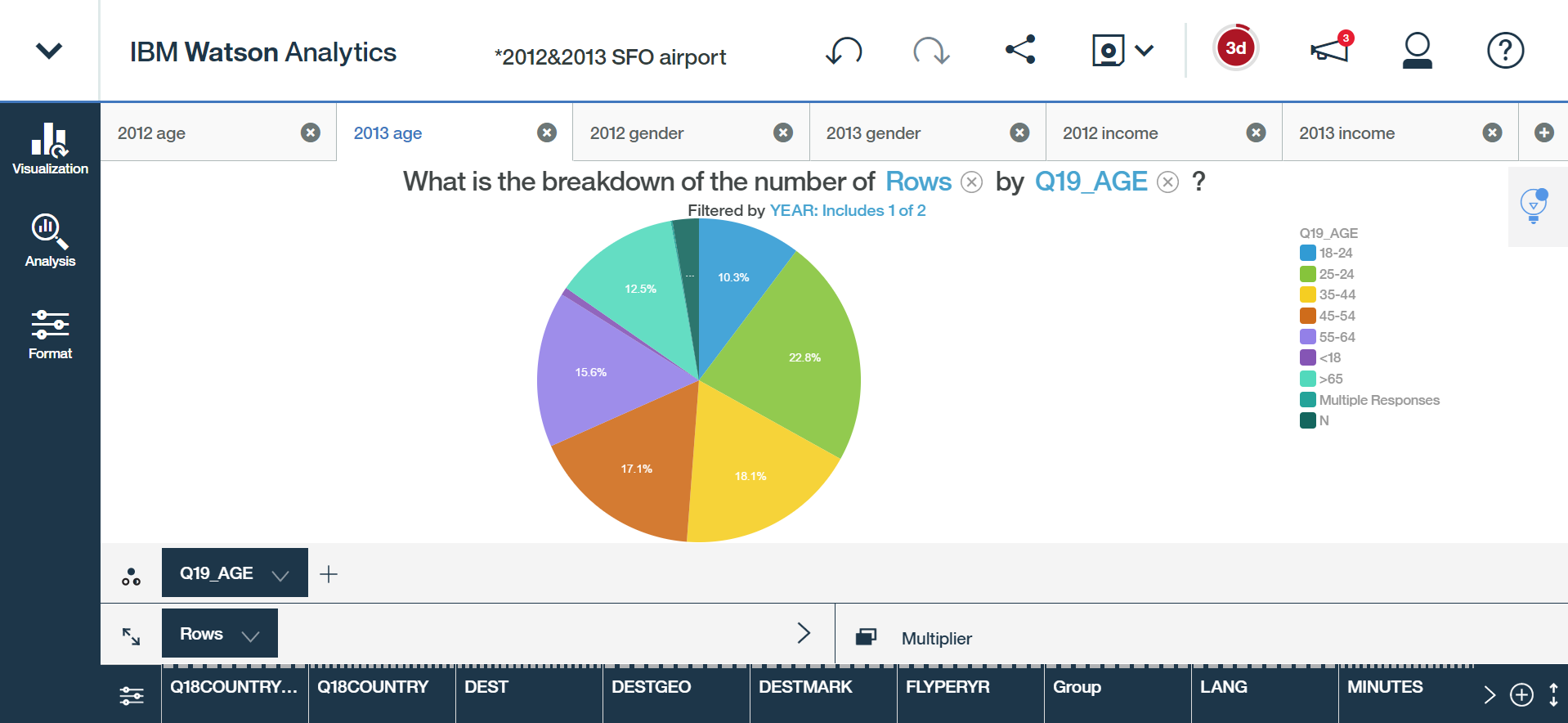
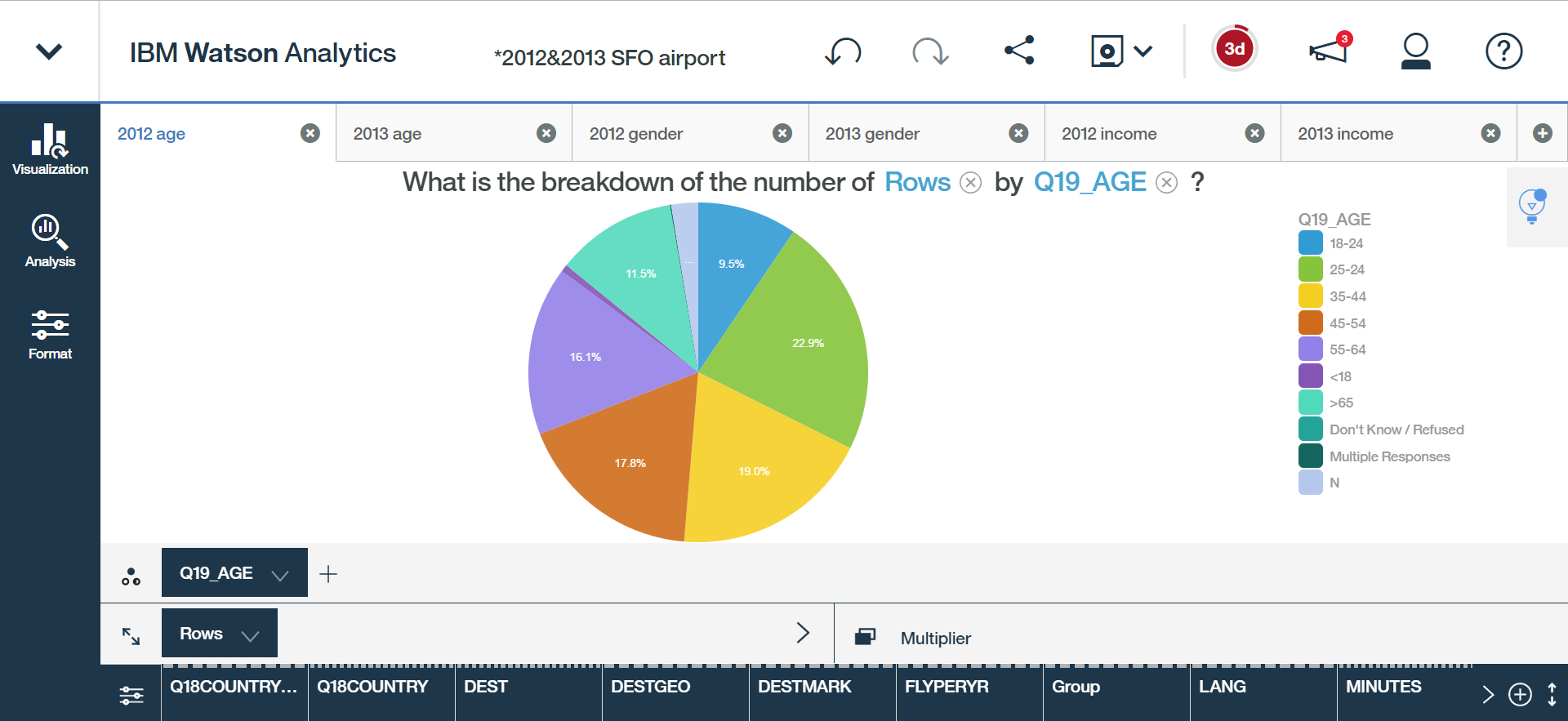
1. **Introduction**

The Commercial Airline Industry is dynamic, complex, and above all else a necessity to a great deal of individuals in today’s fast pace environment. According to the Airports Council International, in 2013 airports generated on average approximately $11.88 in aeronautical revenue and $8.14 in non-aeronautical revenue per passenger. As such, the customer’s experience is of the utmost importance for a thriving and successful airport. The entire customer’s journey, from the arrival at the airport to departing at their destination, is directly impacted by the choices the airport makes. The San Francisco International Airport is no exception.

As a company, SFO must constantly improve their understanding of their customers. Anticipating what the customer wants in the future is best approached by learning from the customer’s previous desires. The 2012 and 2013 Customer Surveys have been of tremendous value to our team of business analysts. We provided to your marketing team an extensive analysis of this survey information. We applied our domain knowledge of SFO and our background in machine learning algorithms in order to segment SFO customers into three distinct groups. Upon grouping the customers, we identified key survey statistics of each group. These statistics are diverse and include elements such as the likelihood of customers using airport services (e.g. parking, check baggage, airport stores, etc.) as well as SFO’s website. The aim of this report is to utilize the above analysis to provide tailored recommendations to SFO that will enhance customer experience and create the foundation for future business strategies.

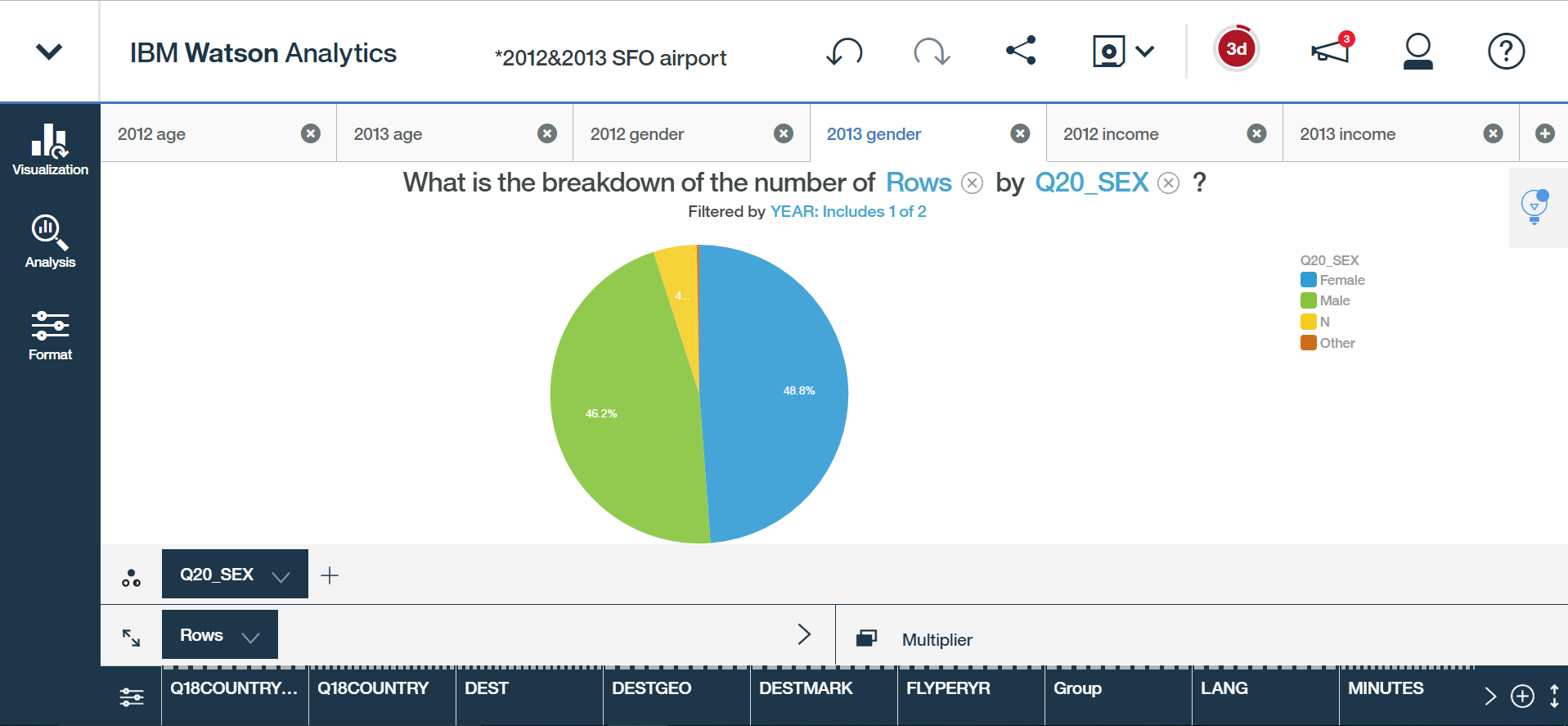
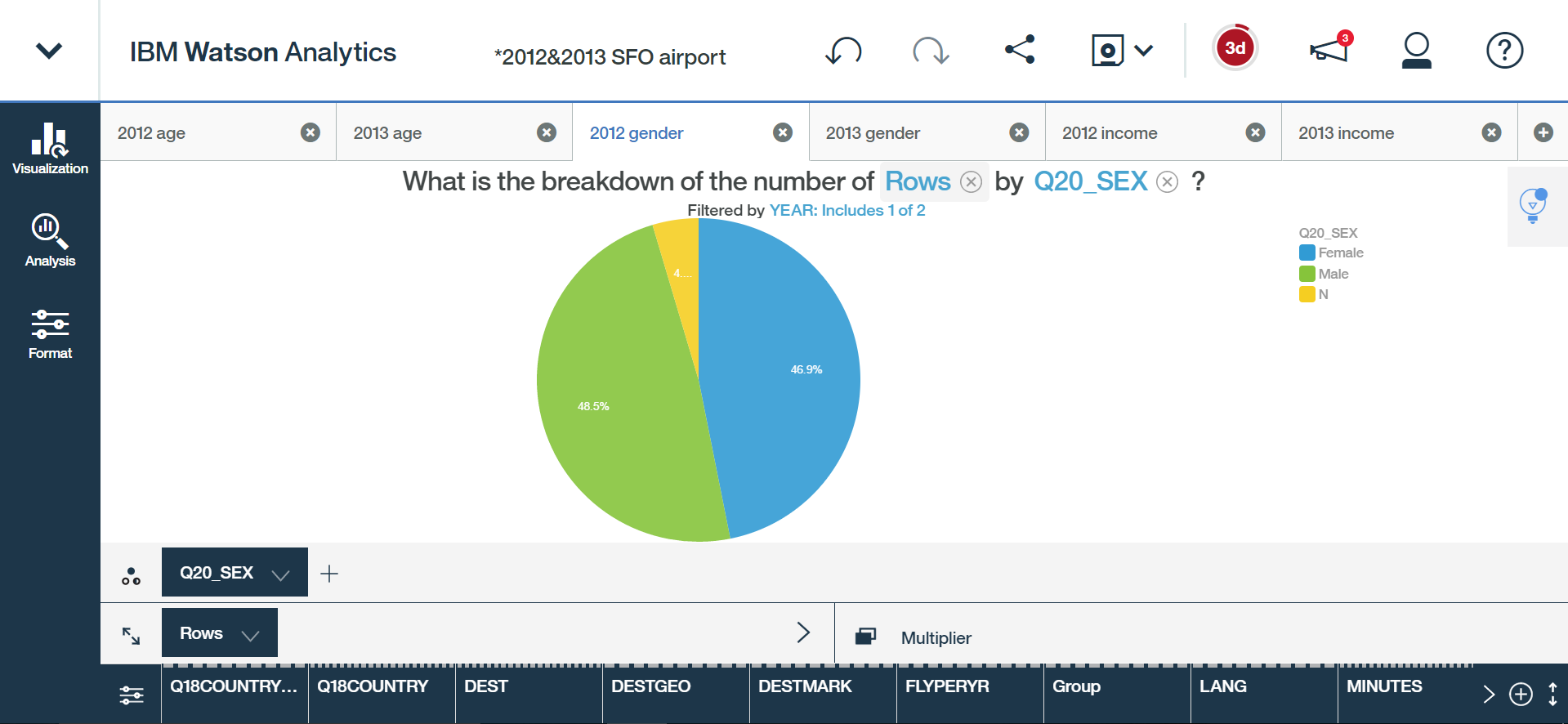
1. **Exploratory Analysis**

Since the amount of survey information provided by SFO between 2012 and 2013 was limited, we first wanted to see if we could merge the two datasets into one. The survey questions between these two years changed; however, some questions remain constant between them. We conduct an exploratory analysis to see if the distributions between the two years, with respect to some common questions, remained relatively constant. Three common attributes of the data were investigated: age, gender, and income level. The following pictures indicate the data about customers in 2012 and 2013.

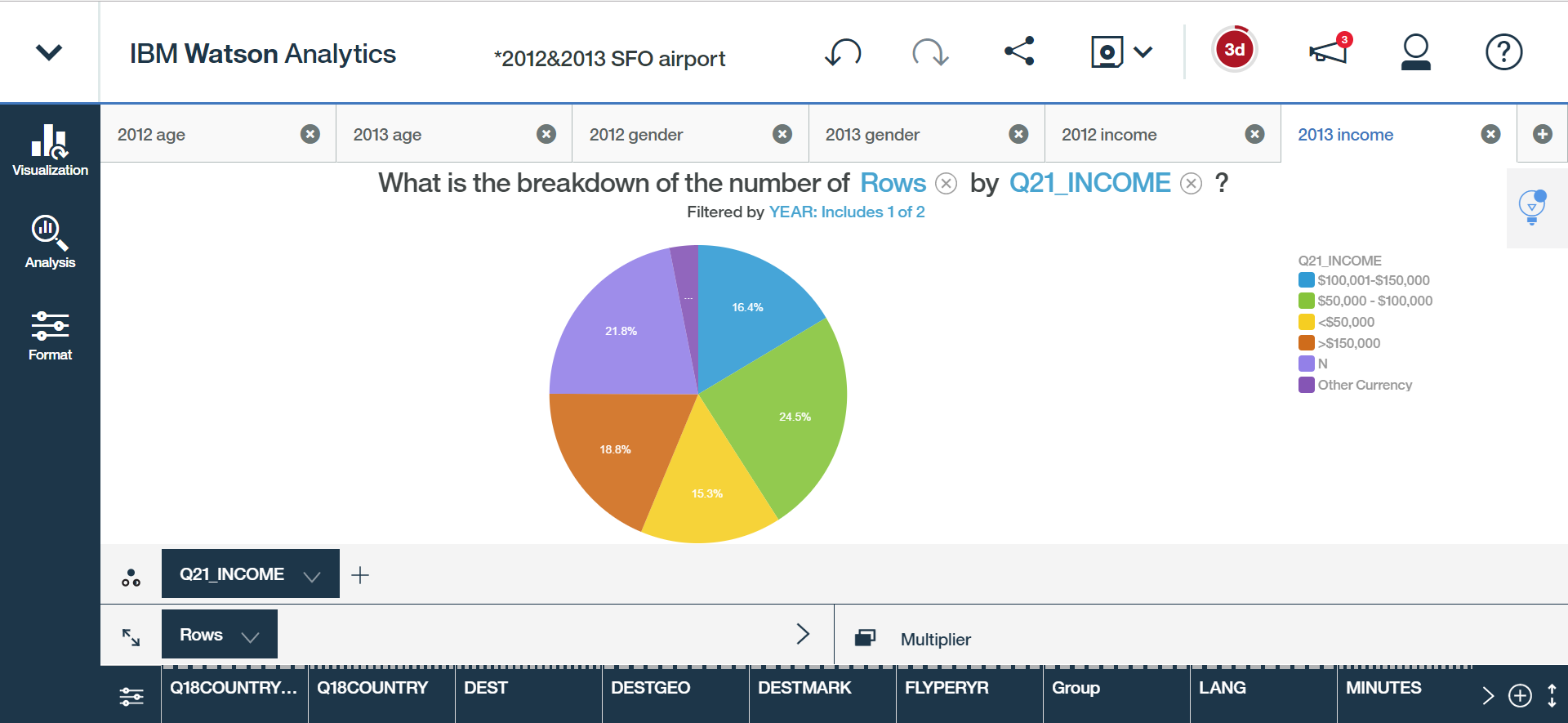
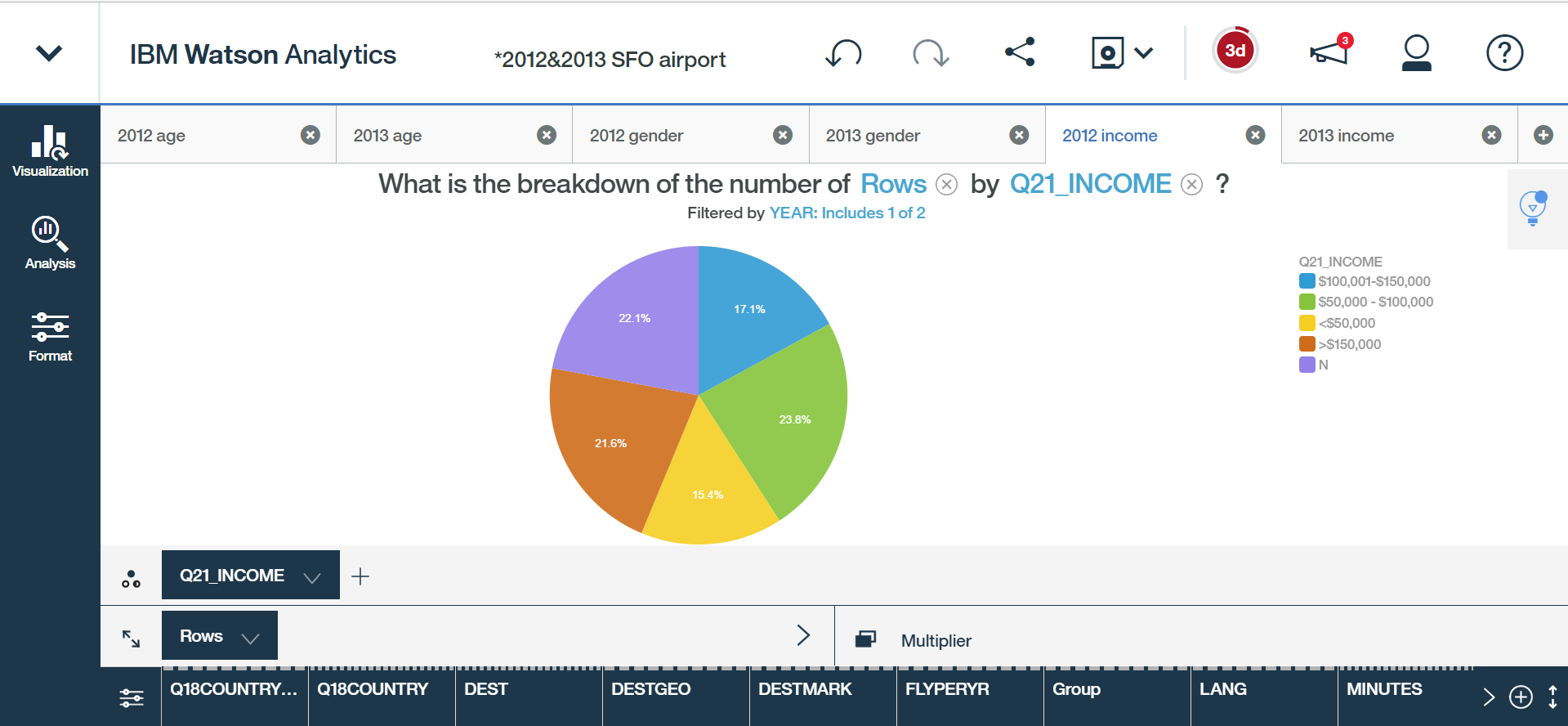


**Figure 1**: The breakdown of age by survey data from the year 2012 and 2013 respectively.

From **Figure 1** can see that the proportion of the main six age segmentation are almost the same. Thus, we say that the customers are in the same age distribution. In other words, the percentage of individuals traveling at SFO between 2012 and 2013 are time invariant.

  
**Figure 2**: The breakdown of sex (gender) by survey data from the year 2012 and 2013 respectively.

**Figure 2** shows that the gender distributions are also time invariant to the year. The deviation is negligible. Customers involved in the survey were nearly half male and half female.

  
**Figure 3**: The breakdown of income by survey data from the year 2012 and 2013 respectively.

When it comes to the income distribution of customers, **Figure 3** shows that the high income customers (>$150,000) seemed to shrink from 21.6% in 2012 to 18.8% in 2013. However, it seems that this difference can be attributed to the ‘Other Currency’ category. This seems to indicate that high income individuals from other countries travel SFO frequently. In all, we can say that customers in 2012 and 2013 has the same income status.

To conclude, we found that there is negligible difference between customers in the survey in 2012 and 2013 with respect to these three major questions. This, however, leads to two outcomes. It could be that SFO is truly surveying individuals at random and those individuals choose to provide their information. In other words, although the questions of the survey changed between the years, SFO is keeping the methodology and distribution of those surveys roughly the same. This could also be due to the limited time frame the sample data was provided. Only the May data is provided for 2012 and 2013. This similar distribution could instead indicate the seasonality of SFO and not the actual general types of customers at SFO at any given time of the year. Thus, we first recommend SFO clearly define their survey methodologies and release more data from different time frames.

1. **Customer Segmentation**

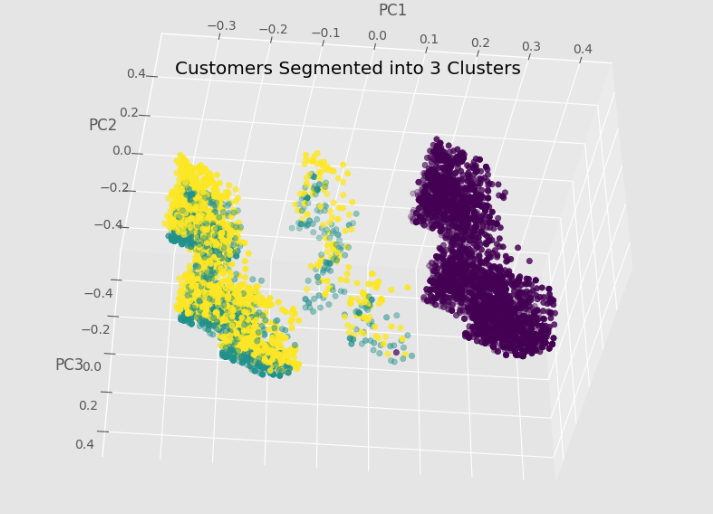
Customer segmentation can be approached from many different ways. One can tackle it from purely a business perspective, utilizing expert domain knowledge to group individuals. One could also tackle the problem from a purely analytic or statistical framework. Our group decided to take on the task of segmenting customers from a mixture of these perspectives.

Essentially, customer segmentation is about dividing a customer base into groups of similar individuals based on specific patterns with respect to a particular market. Customers who share many of the same interests can be grouped together, and business strategies can be specifically tailored to those individuals. The customer surveys provide a plethora of data to compare one customer to another; however, some survey questions are more important or indicative of overall customer behaviors. Of all the survey questions shared between the 2012 and 2013 surveys, we chose eight questions, split amongst three categories, to segment over. The first category is Customer Demographics, and is made up of customer age, gender, country of residence, and household income level. The second is Flight Information, and consists of the market size of the destination airport and the area in the world the destination airport is located (i.e. US West, Europe, Asia, etc.). The final category is Web Usage and is solely comprised of whether or not a customer has ever used the SFO website in the past.

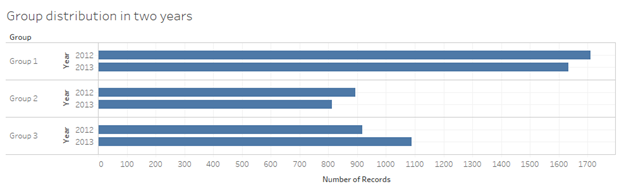
Our team first cleaned and processed all the raw survey data. If a survey question was missing information, we filled in that entry with an ‘N’ (null or empty character). We did not want to discard survey’s needlessly for incompleteness as they still provide usable information. It was also useful to identify areas where several individuals were not providing information specific to a question. The eight questions above were specifically extracted from the data and placed into its own table. This created a dataset of 7,057 surveys (3,522 customers from 2012 and 3,535 customers from 2013).

The responses to each of the combined questions were encoded as ‘dummy variables’ or, in other words, binary variables indicating whether or not each response for a given question was provided. The dimension of the data was then reduced through Principal Component Analysis (PCA). In short, PCA serves to reduce the number of variables that need to be learned from while preserving the majority of the information stored in the additional variables. These new variables, called Principal Components (PC), are linear combination of the original variables. The reason for this was to perform unsupervised *k-means* clustering over the dataset. Briefly, *k-means* is a common cluster analysis algorithm in data mining which aims to partition *n* observations (customer surveys) into *k* clusters (groups). Groups are formed by minimizing the within-cluster sum of squares error distance metric. In this context, customer surveys that are more similar in responses to one group (ex. Group 1) than those in another group (ex. Group 2) are placed in the first group (Group 1). We chose to segment customers into three groups as we believed it struck a balance between differentiating the customers yet keeping the segments consolidated enough for in depth analysis to be conducted on each group. **Figure** 4 contains two figures representing the three clusters in a three dimensional space that let us visualize the survey responses fairly well.

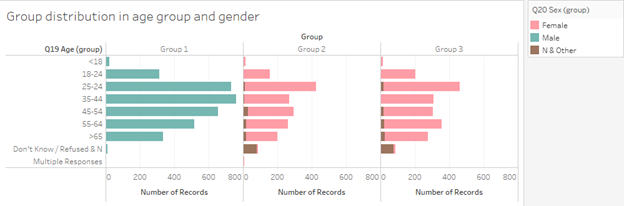
From these depictions, there seems to be more than three groups of customers. This is to be expected. However, the separation amongst the three clusters is quite clear when looked at from multiple perspectives. Since our initial data pre-processing condensed the eight survey questions into Principal Components, it is difficult to see the exact questions that divide these groups. As such, we next conduct analysis to identify the similarities and differences amongst the three customer segment groups.

  
**Figure 4**: Customer survey data shared by the 2012 and 2013 SFO survey questions: country of residence, age, gender, income level, destination area, market size of destination airport, and web usage. The surveys are clustered into three groups. The three dimensions reflect the directions of most variance in the survey data.

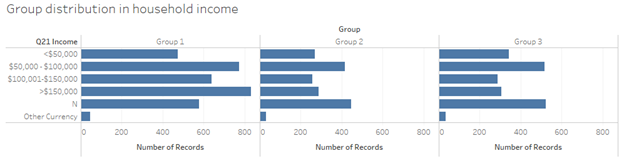
Since we obtain the customer segmentation from a data science perspective, we need to explain how data separated in a plain English as well. First we studied how data is each group disperse in two years.

  
**Figure 5:** Segmented group distribution between the 2012 and 2013 years.

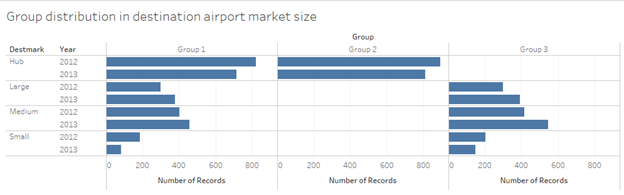
From the above chart we can see each group has close number in both years. Only for group 3, data in two years has a bit larger difference comparing to other two groups. As for year, the three groups doesn’t have obvious distinction. So that we looked at another two factors, age group and gender, to see whether the two groups distinguish clearly.

  
**Figure 6**: Segmented group distribution by age group and gender.

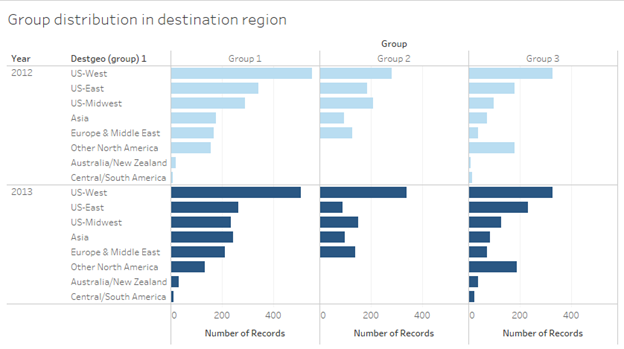
In this part, we can see that the groups have very clear differences - group 1 only has male customers while group 2 and 3 only have female customers and other gender customers. For unknown and other genders, we group them into the same group. So that only 3 gender group is included. The age range, on the other hand, is distributed normally among all groups. A slight difference is group 1 has the age peak at 35-44 while the other two groups have the peak at 25-34. Groups 1 also has a large population in 25-34 though. Age-wise, there are is much difference in groups.

  
**Figure 7**: Segmented group distribution by household income.

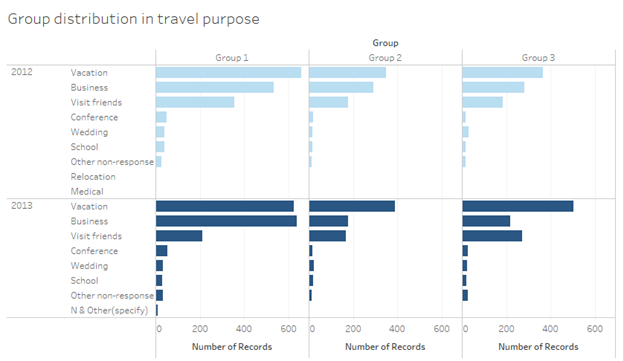
Household income is usually a separator in customer segmentation. So that above is the chart showing differentiation in 3 clustered groups. An obvious difference is that group 1 has more customer with annual income either 50k to 100k or over 150k, while group 2 and 3 have more customer with annual income either 50k or no response. The salary difference may cause various airport service usage answers in other questions.

  
**Figure 8**: Segmented group distribution by destination airport market size.

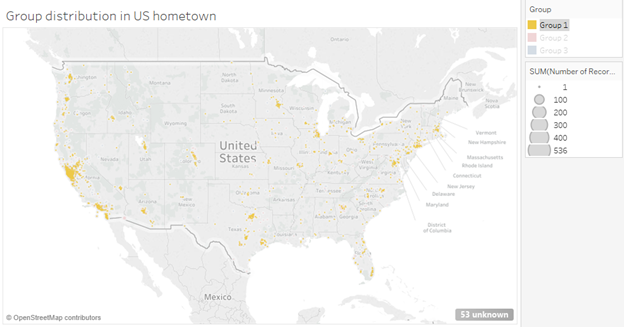
Based on the customer destination airport market size, it is obvious group 1 has covered all market sizes, while group 2 only covers hub size and group 3 only covers large, medium and small sizes. So that we can tell destination is a leading factor in customer segmentation.

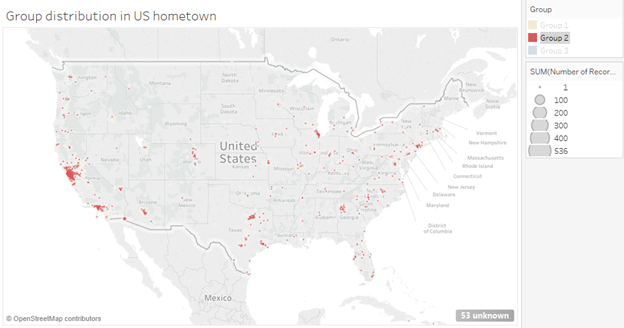
  
 **Figure 9**: Segmented group distribution by destination region and year.

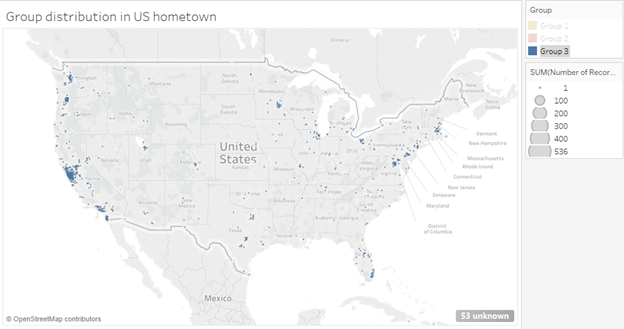
When looking at the destination region in the groups, apparently group 1 and 3 have contained all types of regions while group 2 has contained most of the regions except other North America, Oceania and central/south America.

  
**Figure 10**: Segmented group distribution by travel purpose and year.

Sometimes travel purpose can direct customer’s behavior at the airport, so that the group also study the distribution for clustered groups. It turns out that vacation, business and visit are the top 3 reasons for all groups, the percentile in each group and either year are different though. In 2012, all groups are having most vacation customers, and then business and visit customers. While in 2013, there are same amount people in group 1 doing vacation and business. Also, for group 2 and 3, the number of vacation is more than business and visit combination. This might stimulate distinct answers in the later analysis.





  
**Figure 11**: Segmented group distribution by US only hometowns and year.

**Figure 11** marks the hometown for customers in each segmentation. We can see all groups have crowded dots in the bay area considering the location of SFO. Each group has similar distribution for hometowns so that this might not be an important factor in the customer analysis.

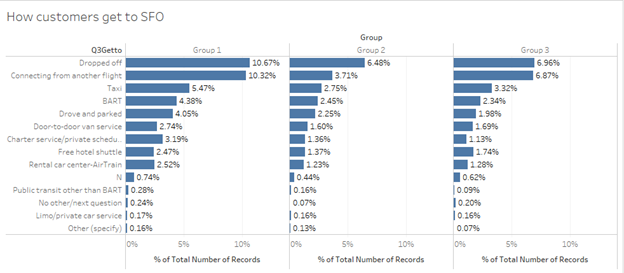
In conclusion, the top features in our customers segmentations are - gender and destination airport market size. Also, destination region and travel purpose plays a role in separating the customers into groups.

1. **Customer Usage of Airport Services**

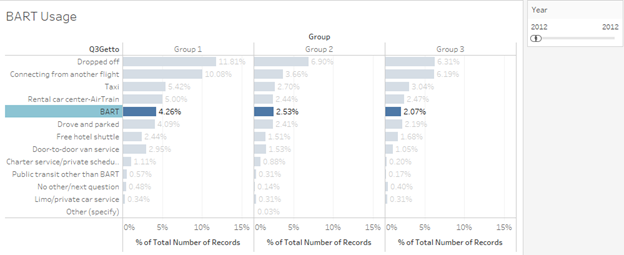
**Transportation**

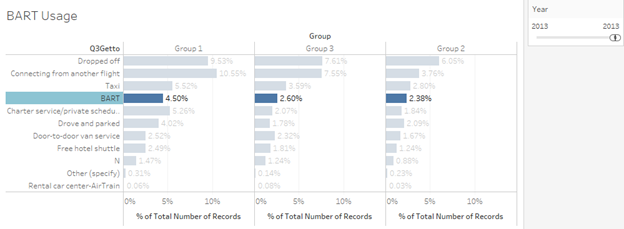
The majority of the two surveys focus on the customers’ attitudes toward various airport services. Based on the customer segmentation above, we sought if differences exists amongst groups regarding their likelihood using airport services and their satisfactory levels.

First we looked at the transportation services for getting to the airport. The following is a chart showing the percentage of customers within each group that used various transportation methods to SFO.

  
**Figure 12**: Segmented group distribution of transportation methods for customers to arrive at SFO.

From **Figure 12** we can see there are majority of customers asking families or friends to drop them off at the airport. This is the easiest and least expensive way. An important selection here is the BART service. BART is the fastest, easiest and least expensive way to get to SFO and to downtown SF. On the SFO official website, BART has its own introduction page in airport connections section. Moreover, BART also helped SFO win a pioneer prize in information technology in the industry. The following two charts show the BART’s percentage in all transportation methods to SFO in 2012 and 2013 over customers who chose to park at SFO. We can see for all groups, BART has gained a little more share. Meanwhile, the purpose of BART aims to release the presser of private or commercial vehicle in an environment-friendly way. Comparing to taxi usage, the percentile of BART is still low. So that our group recommend SFO to market more about BART and educate the advantage of BART. SFO can set up educational board to advertise BART in places like public parking lot, terminal gate and taxi drop-off places.



**Figure 13**: Segmented group distribution by BART usage.

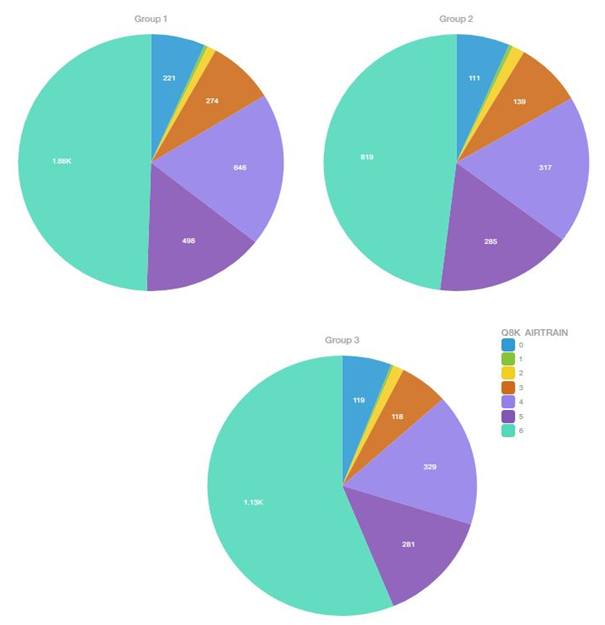
**Parking**

On analyzing the rating of parking facility on the basis of different segment groups, there was not much difference in the result. Further, on analyzing this on the basis of income results were different. Overall it is less dependent on the quality of facility of Airport parking because as the income group increases people have used the parking more. The less income group has more proportion of people who have not used the parking. Thus targeting people with higher income groups for parking would be a good option. Further, about 60% of the people who have used the parking are satisfied and rated 4-5 in the entire income group. It need to be analyzed that why other 40% people are not satisfied with the parking facility to generate more revenue from parking.

  
**Figure 14**: Segmented group distribution by level of satisfaction of airport parking by income.

**AirTrain**

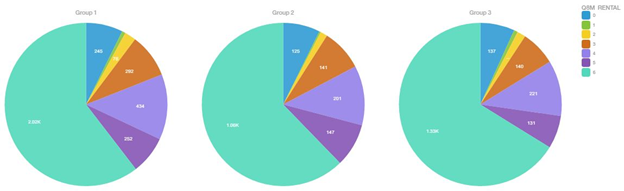
The charts in **Figure 15** show the rating of AirTran of all the three groups. It can easily be inferred from above charts that more proportion of people from group 2 are using AirTrain as compared to other two. Group 3 is the group in which less number of people is using AirTrain as compared with other groups. Although, the satisfactory level from the AirTrain is equal in all the three groups i.e. more than 50% of people have rated the service very good score. Hence group 2 and 1 could be a target audience for AirTrain.

 **Figure 15**: Segmented group distribution by level of satisfaction of airport parking by income.

**Rental Car**

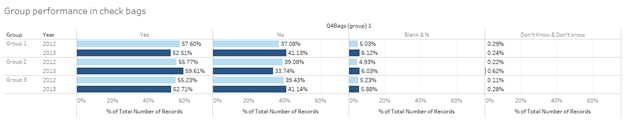
Rental car is one the most used service in an airport. The above pie chart shows the ratings from different groups for the Rental car service of the SFO airport.

As we can see, group 1 uses the rental car most and more than 50% reviews are positive in all the groups. The potential and most lucrative audience would be of group 3 than group 2 and at last group 3. The marketing activity implemented in this order would lead good results.

  
**Figure 16**: Segmented group distribution by satisfaction level of rental car services.

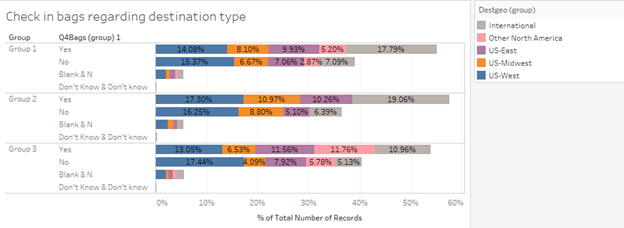
**Check Bags**

After customers arrive SFO, the first thing they will do obtain a boarding pass at check-in table or kiosk if they did not utilize a virtual boarding pass. Meanwhile, they will determine whether they check bags or not. The following chart shows the check bag results in each group. We can see each group has very close performance in check bags between 2012 and 2013. For group 1 and 3, the percentage of “check bags” is decreasing on a small rate, while for group 2, this percentage increasing a little bit.



**Figure 17**: Segmented group distribution by whether or not customer checked bags.

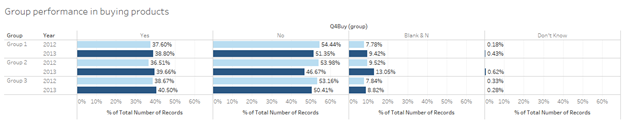
Since we want to have deep understanding of how performance is related to other factors, the following chart shows the relationship between check-in baggage and destination type. The percentile indicates how each destination performs within the customer group.

  
**Figure 18**: Segmented group distribution by whether or not customer checked bags and destination.

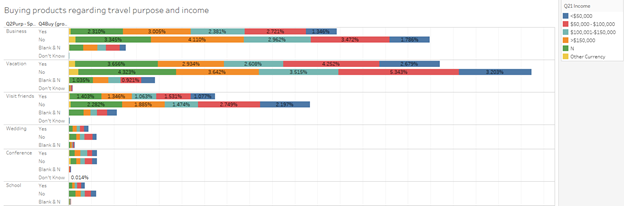
The check-in baggage performance differs group-wise. For all the “check-in” customers, in group 1 and 2 there are more traveling internationally comparing to national-wide while group 3 has more US-west travelers check in bags. On the other hand, group 2 has more US-midwest travelers who are not checking in bags comparing to other two groups. Both of the difference is small but existing.

**Store Purchase**

Store purchases can be made anywhere inside the terminal. It is also an important incoming factor for the airport. When we analyze the purchasing behavior among all groups, it shows that the likelihood for them to buy products in the airport if very close. The data has also shown a slight increase in buying from 2012 to 2013. The following chart has provided visualization support.

 **Figure 19**: Segmented group distribution by whether or not they bought a product in the airport.

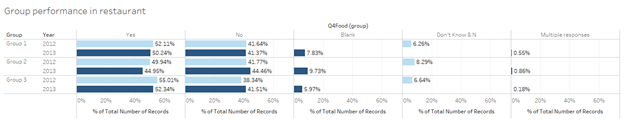
However, there can be other factors affecting people’s buying. So that we take a deep analysis in the relationship of purchasing between customers’ income and purpose of visit. The following chart has shown the result.

  
**Figure 20**: Segmented group distribution by whether or not they bought a product in the airport and income level.

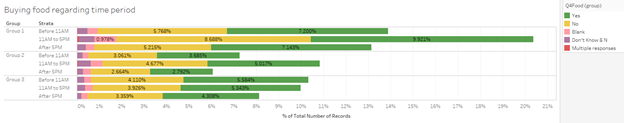
The graph in **Figure 20** shows information about proportion of people who buy something from airport stores. These people are classified on the basis of their purpose of visit and Income. It can be noted that, there are more customers who are not intended purchase goods at the airport comparing who do irrespective of their purpose of visit. Amongst the “buy” group, vacation has the highest number. Also, income difference doesn’t differ much in buy or not.

**Restaurant Consumption**

Food services is another key part in the airport. According to the following chart, customers in all groups are tending to buy less food from airport restaurant. Such reduction should be relate to the actual restaurant sales to see whether it is true in the whole airport. It might due to the small sample size of the survey also.

  
**Figure 21**: Segmented group distribution by whether or not they purchased food.

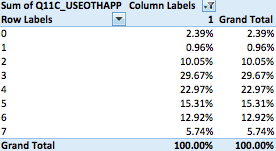
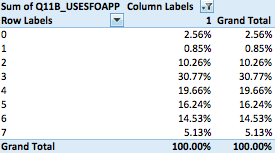
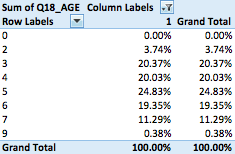
Since major food purchase is very related to the time, we have developed the chart to show customer’s food purchasing in different time period. The result is shown as below.

  
**Figure 22**: Segmented group distribution by whether or not they purchased food by time period.

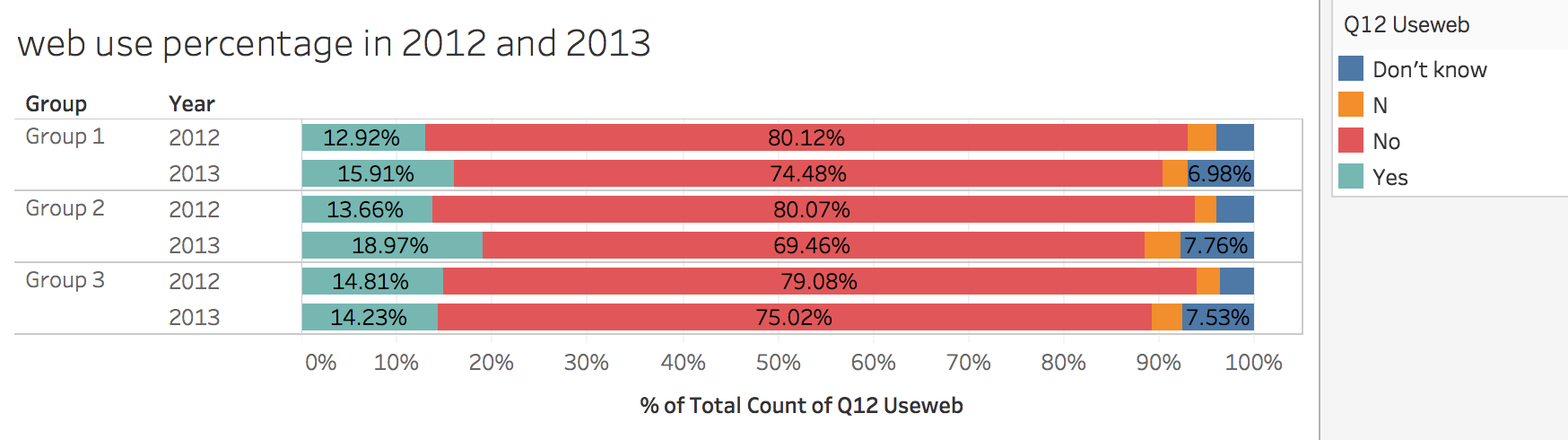
Based on the time period, we can see the groups have different performance. For both group 1 and 2, their purchasing rate during 11AM to 5PM is higher than other time, while group 3 has very similar food buying performance during the three periods. A good sign here is, if we see the entire data set as a whole regardless year, there are always more “buying” customer comparing to the “no buying”. So that the key is to transfer the “no buying” to buying. Also, for people in group 3 who travels between 11AM to 5PM, since group 3 are women customers who are going to non-hub airport, they might prepare own food so that the difference amongst three periods are very close.

1. **Customer Usage of SFO Website and Related Amenities**

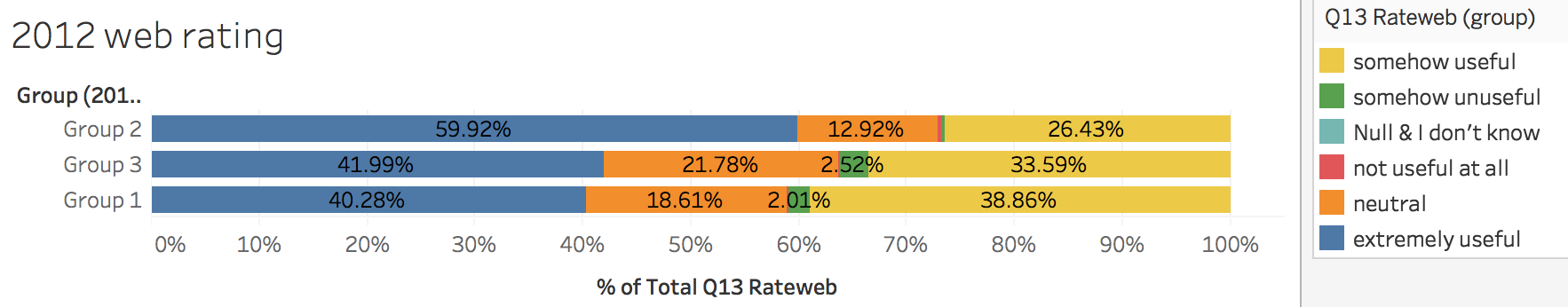
From **Figure 22**, we can see that customers who use website are mostly from age 25 to 64, age 25-34 are mostly the people who are using the application, age 25- 44 are people who are using the other applications most. From above analysis, we can see that age does differentiate people who use website more often or not. It appears that people from age 25 to 34 are the majority of people who would browse the website more often.



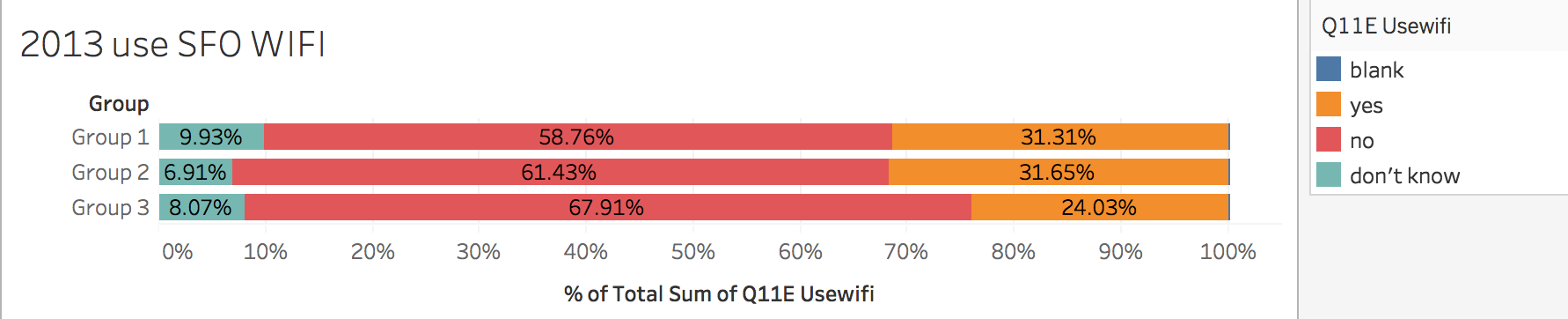
**Figure 22**: Basic distribution of age groups who use the SFO website between 2012 and 2013.

  
**Figure 23**: Segmented distribution of groups who use the SFO website by year.

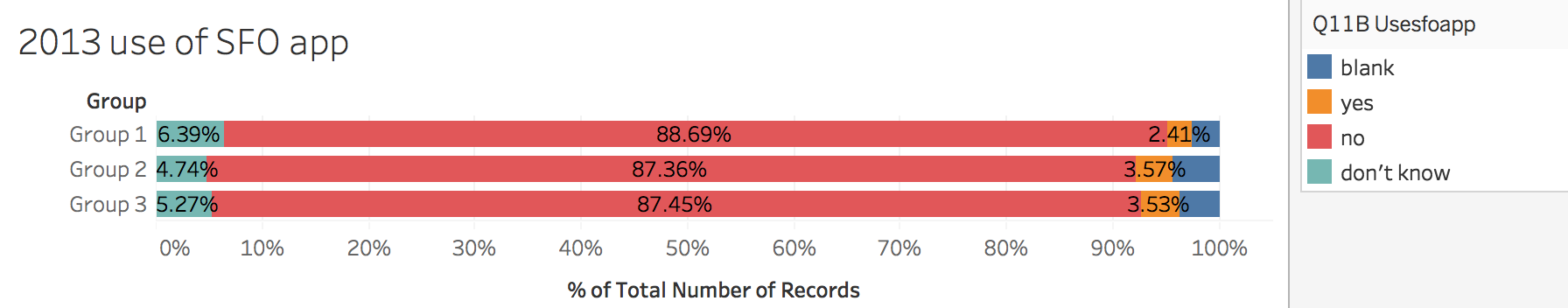
The above chart shows the website usage status of each group in 2012 and 2013. Though a large proportion of people don’t use the SFO website, we can see that people who use SFO website have increased in both group 1 and group 2 from 2012 to 2013. Group 2 has the largest increasing from 2012 to 2013. In group 3, though there is slightly decreasing for people who use the SFO website, the percentage of people who don’t use the website also decreased. The overall trend of SFO website usage is that more and more people are trying to use the SFO website.

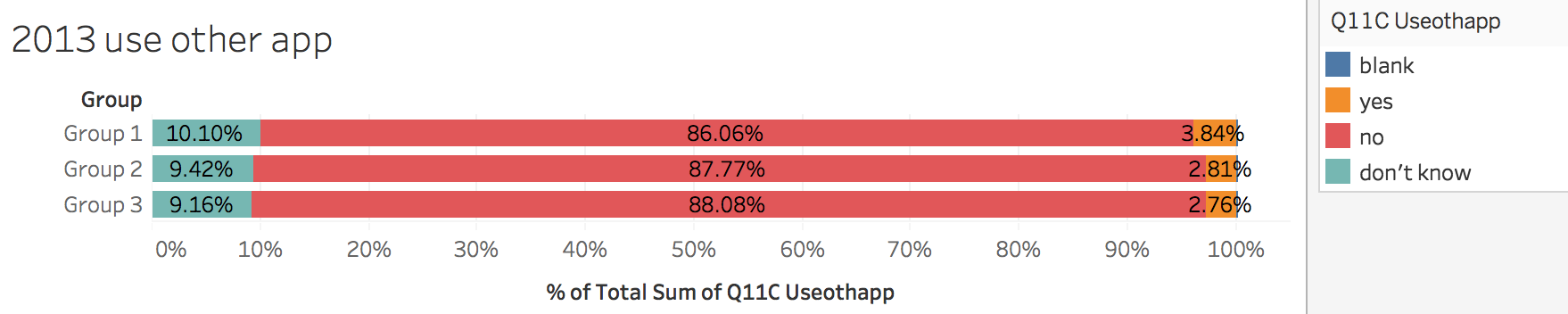
  
**Figure 24**: Segmented distribution of groups used and rated the SFO website in 2012.

**Figure 24** depicts the analysis about how people feel about the SFO website. From the analysis we can know that most of people think that SFO website is useful or at least neutral. Only few people think it’s un-useful. Among the three groups, group 2 has the largest percentage of people who think the SFO website is extremely useful.

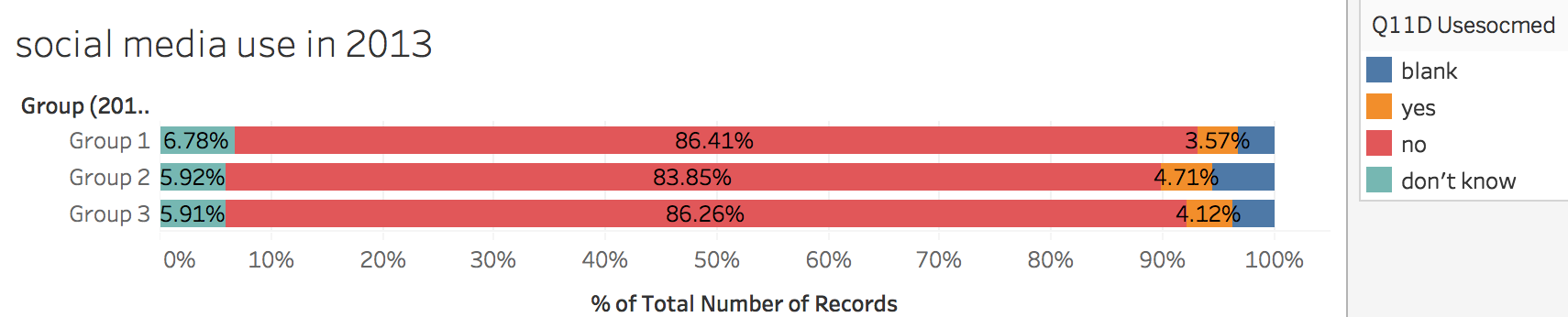
  
**Figure 25**: Segmented distribution of groups usage of SFO Wi-Fi in 2012.

The above chart shows the proportion of people who use the SFO WIFI in each group. We can see that both group 1 and group 2 have the similar percentages, which are about 31%, of people who use the SFO WIFI. However, there is only about 24% people in group 3 that use the SFO WIFI.



  
**Figure 25**: Segmented distribution of groups usage of the SFO mobile app and other similar airport mobile apps in 2013.

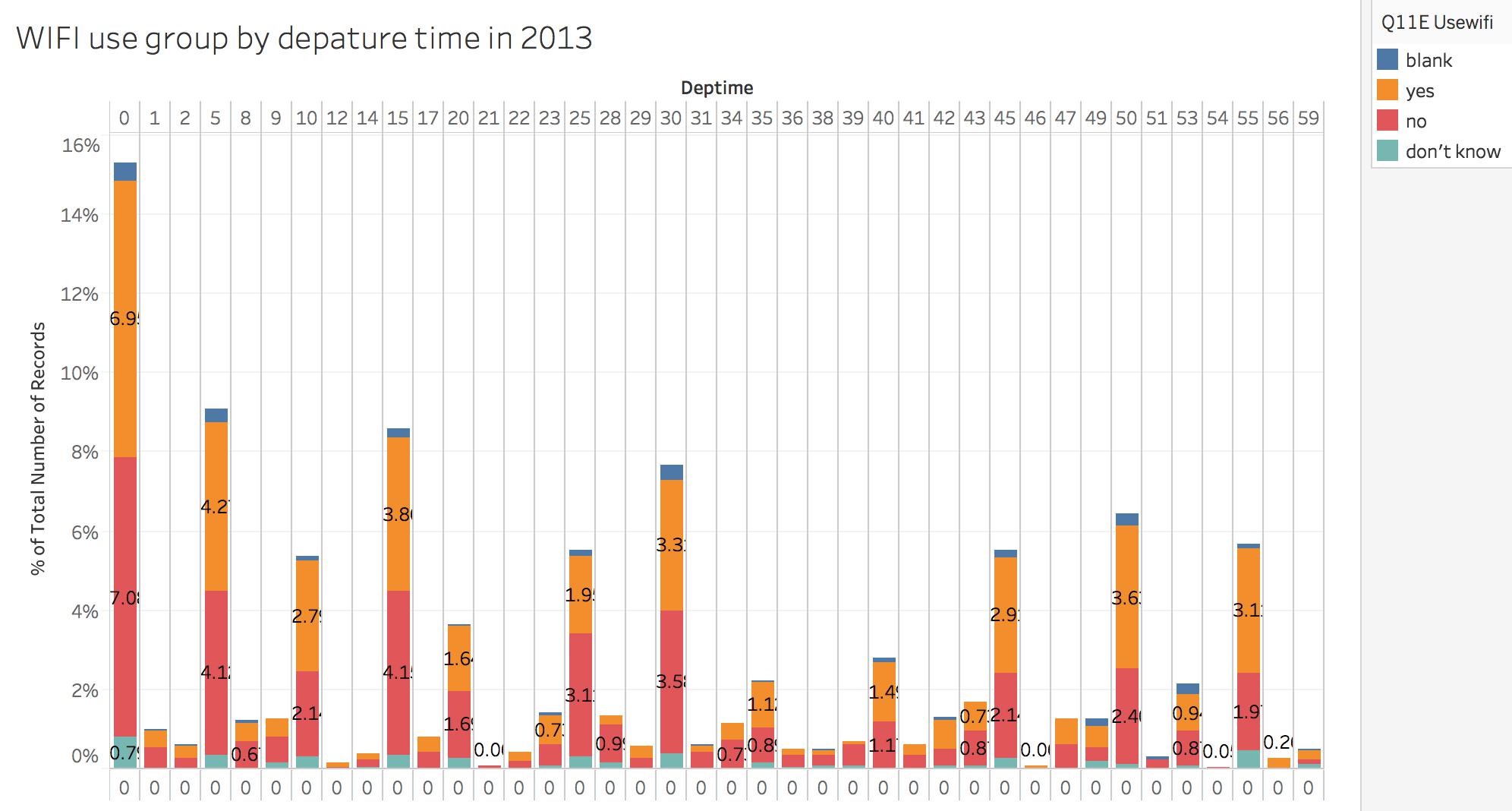
The above two chart describe the usage of SFO application and other applications. There isn’t much difference among the three groups in each chart. Also, the percentages of people who use the SFO app are much similar with the percentages of people who use other app. Only about 3% of people in each group use the SFO app or other app. Maybe because few people are familiar to use those app products that related to airline industry.

  
**Figure 26**: Segmented distribution of groups usage of social media in 2013.

We analyzed the usage of social media in 2013 among the three groups. There isn’t big differences among them. All the three groups have a small percentage of people who use the social media.

Based on the information provided by all above charts that related to web service. We can see that the overall usage of web-related services are not very much. Only few people use the SFO web technology related services in 2012 and 2013. However, the rating of the SFO website in 2012 is pretty good. Most people think that the SFO website is useful. Also, the percentage of people who use SFO website increased in 2013. Thus, the reason that people don’t use the website may just simply because they don’t know there is a SFO website. So it’s better for SFO to advertise their website to customers. For example, SFO can cooperate with some flight booking system or airline companies. They can email SFO website link and advertisements with the booking information to customers whose departure location or destination is SFO. Thus the SFO website will be known by more people. The next time, when people want to check the flight schedules, they may use the website. The survey of 2013 adds several questions, which are about apps and social media, that are not included in 2012 survey. The percentages of people who use these media are pretty small and much similar, which are about 2%-4%. However, comparing to the percentage of people that use the SFO website and apps, much more percentage of people use the SFO WIFI. Maybe because the WIFI is more meet customer’s need, comparing to the apps and website.

Among the three groups, group 2 has the largest increase from 2012 to 2013 and has the most positive attitude toward the SFO website. Also, one third, which is a large portion, of people in group 2 use the SFO WIFI. Thus we can conclude that this group of people perhaps are easy to be attracted by technology services and easy to accept new technologies. However, people in group 3 that use the SFO website barely increased from 2012 to 2013. Also, only 24% people in group 3 use SFO WIFI. It means people in group 3 may don’t need the WIFI or technology-related services as much as people in the other two groups. Thus it may be a little difficult to convince people in group 3 to use technical related services.

  
**Figure 25**: Segmented distribution of groups usage of the SFO WiFi by departures time in 2013.

1. **Recommendations**

Our analyst team would like to end this analysis by providing some general recommendations to the SFO marketing department. First of all, we believe this department is doing an excellent job at SFO. The majority of survey information was completed. The questions asked described a multitude of different aspects of the customers. In addition, content was well updated from year to year based on new aspects SFO has incorporated into their business model.

As the BART services is a creation product in SFO, we recommend increased emphasis on how to increase the usage of BART. SFO can set up an educational board to advertise BART in places like the public parking lots, terminal gates, and taxi drop-off zones.

Although we believe the surveys have been updated from year to year, more effort should be made to make each year’s data comparable to the previous year. We were limited by the amount of questions that could be compared from 2012 to 2013, even excluding the questions which seemed to be directly stemmed from previous responses. This would make data analysis more efficient and less prone to manual errors. Automated data analysis could even be conducted if the data was collected, stored, and transferred in a standard way.

In addition, we recommend continuing the marketing campaign in all types of airport stores. Between the years, the analysis showed that there was a small increase in percent of airport purchasing of products. Although small, this is not negligible and may be able to be compounded upon in future years. However, the data has shown that airport restaurant performance is decreasing ever so slightly. Restaurant performance should remain constant or increase each year. Restaurants can launch promotion meals or special combos during peak period requests. During peak food hours, these promotions should become more visible by customers such as on strategically placed TV monitors. On the other hand, airport can post online or offline surveys on what type of food or price range the customers want in the airport so to increase relative revenue for each customer segment.

Finally, the people who use the SFO website can be seen to increase from 2012 to 2013. Whether this is due to the popularity of the SFO website itself, the advancements of the marketing team, or solely the movement in the United States toward smart phones and other personal information technologies, is unclear. SFO may cooperate with airline companies or some ticket booking systems. The marketing team could potentially send the SFO link out or force users to enter the website from their personal WiFi.