NTNU
Kunnskap for en bedre verden

Department of Computer Science

TDT4259 - Applied Data Science

# Telecommunication customer loyalty analysis

*Authors:*
Martí Costa (568984, martco@ntnu.no)
Miguel Lourenço (566524, miguelab@ntnu.no)
Rafael Lourenço (566569, rafaelam@ntnu.no)
Erling Olweus (489937, erlingfo@ntnu.no)
Jonas Klepper Rødningen (474194, jonasrod@ntnu.no)
S. Xiao Fernández (567214, sofiaxf@ntnu.no)
Group 42

April, 2021

# Table of Contents

# List of Figures

# List of Tables

# 1 Introduction

## 1.1 Verizon Wireless

Verizon Communication Inc. is a multinational telecommunication conglomerate headquartered in New York. It was born as the result of the break up of the AT&T Corporation's Bell System into seven companies. As for today, Verizon is the second largest telecommunication company after AT&T and its mobile network is the largest wireless carrier on United States with more than 120 million subscribers. This branch of the Verizon Communication Inc, known as Verizon Wireless, offers a variety of mobile phone, home phone, internet solutions and services all around the country and claim to provide access to their 4G LTE[1] to 95.9% of the US population.

Their whole business model (Figure 2) is based on their customers and the quality of the company's products and services. This means that a good understanding of their customers' needs and desires, habits of consumption, and other aspects as demographics or social-economical context is crucial for the corporation to guarantee the minimum churn rate[2] and maximize a long term profitability. For this reason we have decided to look into customer demographics in this project, with the intention of eventually obtaining recommendations that could lead to a reduction in churn rate.

As for this report, we will assume Verizon is our client and therefore refer to it as so.

## 1.2 Business Problem Definition

Telecommunication companies tend to have similar services. They usually offer internet services, phone services, among others, which only vary in specific values such as the amount of minutes per month a customer can spend on voice calls. Because of the similarities in the services, telco customers find themselves often choosing between companies every few months, which causes a high churn rate in any telco company. Verizon Wireless is no different (Figure 21). With so many other companies offering similar services, it's easy to lose customers because other companies have better values for their services. As such, the business problem we'll be tackling is Verizon's high churn rate.

Considering the dataset we have, which is explained with more detail in the Dataset description (Section 3.1), and that our client has a high ratio of churning customers, our intention is to use the data analysis to reach recommendations that will lead to a reduction of the churn rate.

In order to do so, we first came up with several use cases that were later mapped on a feasibility matrix (Figure 3) that helped us to decide which of the use cases are more attractive, finding a balance between business value and analysis feasibility. (Schmarzo, 2019)

In short, the data analysis in this project is focused on our client's customers and we will be using a dataset that contains information about their demographics and interactions with the company in order to identify loyal customers' profiles. Once we have identified them, we intend to use this information to make recommendations that will have an impact on reducing the churn rate by, in theory, increasing the number of loyal customers.

## 1.3 Group Members

This group consists of six students with different academic backgrounds and experiences that are summarized in the Table 1.

[1] 4G LTE stands for Long Term Evolution and it is a standard for wireless broadband communication for data terminal and mobile devices. It is the previous technology to the 5G.

[2] Churn Rate refers to the amount of customers that leave the company for whatever reason.

| Members | Background |
|---|---|
| Martí Costa | Martí is in his 4th year of a Bachelor's degree in Electronics Engineering and Telecommunications in Spain and wanted to widen up his knowledge and learn about new fields such as data analysis. |
| Miguel Lourenço | Miguel has a Bachelor's degree in Computer Science and Engineering. He is looking forward to learn more about machine learning and data analysis. |
| Rafael Lourenço | Rafael is in his 1st year of Masters in Software Engineering, specializing in Artificial Intelligence. He chose Applied Data Science to better understand the process of data analysis. |
| Erling Olweus | Erling is a 4th year computer science student doing a masters in software systems and have also worked with artificial intelligence in a startup. |
| Jonas K. Rødningen | Jonas has a bachelor's in informatics and is now doing a master on Artificial Intelligence. He has also worked in several machine learning projects. |
| S. Xiao Fernández | Xiao is in her 4th year of Bachelor's degree in Computer Science and Engineering. She took this course as she is interested on artificial intelligence and machine learning. |

Table 1: Group members' academic background

## 1.4 Roles and Responsibilities

Taking our academic backgrounds and capabilities into consideration, we assigned tasks to each member according to our strengths. The majority of the data analysis was assigned to Erling, Jonas, Miguel and Rafael while Xiao and Martí were responsible for the business point of view of the project.

All the main tasks and their organizations are stated in the Table 2

| Task | Description | Members assigned |
|---|---|---|
| Analysis opportunities | Understanding the company's context and business model, finding a data-analysis problem and uses cases that matched the data set. | Martí and Xiao. |
| Data description | First exploration and description of the data set to find relevant features to our use case. We used Deepnote[3] to plot and map the data. | Rafael and Miguel. |
| Data verification | Analyze the consistency and the usability of the data. | Erling |
| Data preparation | Selecting the most relevant features, cleaning the data, generating attributes. | Rafael and Jonas |
| Analysis | The most relevant tasks were defining "a loyal customer"and coming up with a type of analysis such as making a geographical representation of the most loyal customers and presenting the ratio of loyal customers. | Erling and Miguel |
| Discussion and implications for business | Describe recommendations on the use cases according to the data analysis results. | Martí and Xiao |

Table 2: Task organization

More detailed information about the data preparation, the model and the analysis can be found on the Method chapter (Section 3).

# 2 Background

## 2.1 Objectives

The business objective we chose to support was to reduce the churn rate by increasing the number of loyal customers.

After carefully analyzing the data, and discussing the best approach for our problem by looking at the feasibility matrix in Figure 3, we decided our data analysis use case would be: *Where do the most loyal customers come from?*

Here is a high-level overview of our approach:

- Define the term *loyal* customer in this context.
- Do necessary pre-processing of data.
- Plot the customers per respective county.
- Analyze the results and formulate a scoring metric to establish our recommendations.

### 2.1.1 Data Use Case

We began by doing some exploratory data analysis and verifying the data quality.

Accordingly, we tried to understand what features had more value given our objective of reducing churn rate, though we had not decided on an analysis use case yet. In order to have a better understanding of the data we plotted several graphs to visualize individual features and pairs of features (these are presented later in Section 3.2).

Additionally, we looked for correlations between features, to gauge the potential of exploring the creation of predictive models related to churn.

We used our insights from the exploratory data analysis as fuel for coming up with possible data use cases. As an example, we used the plot in Figure 1 to understand the pros and cons of Verizon and what exactly its customers are not happy with.

The use cases will be explained later in the report in the Data Strategy subsection (Section 2.2).

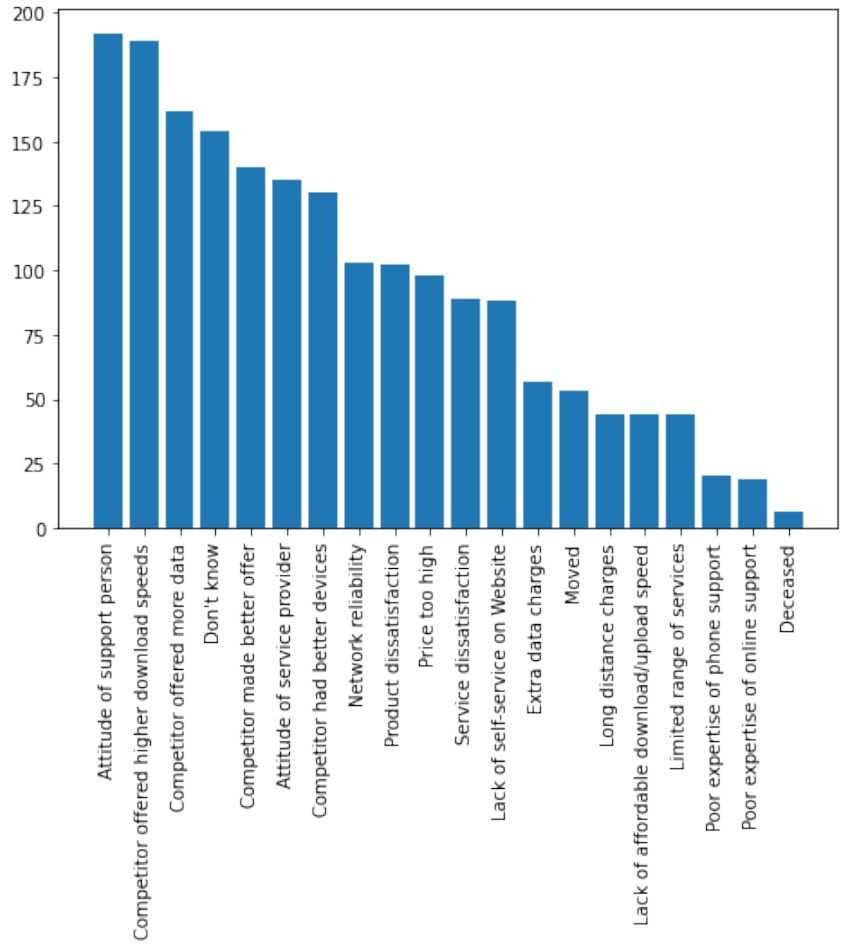


Figure 1: Reasons to churn

This plot is a very good indicator of how we can tackle our objective and try to reduce churn. Because we are able to define a starting point and understand exactly why Verizon's customers are defecting. Through knowing why they churn, it gets easier to look for potential solutions. By looking at Figure 1, we see that out of those who churned, Verizon is losing most of its customers to the competitors.

With all these in mind, we created some use cases that could help achieve our objective:

1. **Where do Verizon's most loyal customers come from?**

   This will help Verizon understand its strengths as a company. If the company is conscious about where their loyal customers come from, they can for instance increase the advertisements in areas with higher loyalty. This can be achieved through making a geographical segmentation of customers. Here we assume that the chances are higher that a new customer from a "loyal area" is more likely to be loyal. Understanding the demographics of the loyal areas could also lead to a deeper understanding of what customer segments, in general, are likely to be loyal with Verizon's current products. If Verizon gets more loyal customers, their churn rate will improve, which is our objective.

2. **Figure out what products our users are not satisfied with.**

   We will figure out what better offers our client could provide to its customers to make them stay or to make them choose Verizon among others. The company will also be able to improve the products that the customers are not satisfied with.

3. **Who has the highest churn rate?**

With this use case we can identify the different customers profiles and focus on those with higher churn rate. This will let us know which social groups have higher churn rate and therefore our client should focus its offers and advertisement on. Hence, preventing customers from churning and convincing them to stay with Verizon for longer.

4. **Find out which customers are more loyal to the company.**

   This will let our client know which characteristics are the most common among all its customers and will help them know where to focus our advertisements. A study of the viability of releasing new plans and offers to attract customers with those characteristics can be drawn from this use case.

5. **Who is more likely to make referrals.**

   With the information obtained from this use case, our client can make a study and try to incentivize customers to bring family or friends to the company with special offers for the customer that made the referral and the new customer that was referred.

   Making referrals and recommending Verizon services is also a good way to create social proof, which in turn may increase the loyalty of the customers.

## 2.2 Data Strategy

We used a combination of components from well-known data science project management techniques. Specifically: Cross-Industry Standard Process for Data Mining (CRISP-DM)[4] to help us plan, organize, implement and iterate our solution (Wirth and Hipp, 2000). Additionally, we also used Business Analytic Model (BAM)[5] to help us understand the business, structure the problem and to generate and choose use cases (Hindle and Vidgen, 2018).

These methodologies provide some important guidelines that we should follow in order to make the best informed approach to designing our data strategy. The main concepts of CRISP-DM are:

1. **Business Understanding**: Figure out the objectives and requirements of the project. In this step we chose to use BAM to help better understand the business context.
2. **Data Understanding**: Identify, collect and analyze the data sets that can help achieve the project goals.
3. **Data Preparation**: Prepare the data set for modeling.
4. **Modeling**: Build and assess various models based on several different modeling techniques/angles. We approached the use case in different ways through introducing two different definitions of loyalty.
5. **Evaluation**: Decide which model best meets the business goals.
6. **Deployment**: Present the results in a useful and understandable manner.

For resolving the best use cases to help reduce the churn rate, we followed BAM strategy (specifically stages one and two), starting by mapping the business model and doing the canvas, shown in Figure 2.

[4]CRISP-DM is an industry-proven way to guide data mining efforts. It includes descriptions of a project's phases, the tasks involved with each phase, and an explanation of the relationships between these tasks.

[5]BAM is a methodology used to understand a company and its context to better describe the business initiative, how it affects to the mentioned company and which use cases can drive to the desired result.

<table>
<tr><th>Key Partners</th><th>Key activities</th><th>Value propositions</th><th>Customer Relations</th><th>Customer Segments</th></tr>
<tr><td rowspan="3">Devices and accessories resellers<br>Companies that join Verizon Innovation Program<br>Device manufacturers<br>Alcatel-Lucent<br>Cisco<br>Ericsson<br>Qualcomm<br>Samsung</td><td>Services design and development<br>Sells of devices and accessories</td><td rowspan="3">High performance<br>Innovation<br>Brand & status</td><td>Support section on the website with Q&A.<br>Virtual chat Assistant<br>Internet forum<br>Phone support<br>in-store workshops</td><td rowspan="3">Consumers<br>Business<br>Carriers<br>Government</td></tr>
<tr><td>Key resources</td><td>Channels</td></tr>
<tr><td>Innovation Centers<br>Industry experts<br>Sales/marketing employees<br>Customer service staff</td><td>Retail stores network<br>Website<br>Telemarketing sales department<br>External chains (Wall-Mart, Best Buy, etc).<br>TV, Radio, Internet, point-of-sale channels</td></tr>
<tr><td colspan="3">Cost structure</td><td colspan="2">Revenure streams</td></tr>
<tr><td colspan="3">Selling & Administrative expenses<br>Cost of services<br>Wireless cost of equipment</td><td colspan="2">Product sales<br>Subscription sales<br>Advertising Sales</td></tr>
</table>

Figure 2: Canvas

After finishing the business canvas we created the matrix from the use cases we had, making a data driven decision about which of the use cases is more worthy, taking into account the business value and data analysis feasibility.

From the canvas, Figure 2, we arrived to the conclusion that the new service wishes to make Verizon a more high performance and innovated-oriented enterprise by offering support and more assistance to the customer. With that information in mind, we did the feasibility matrix.

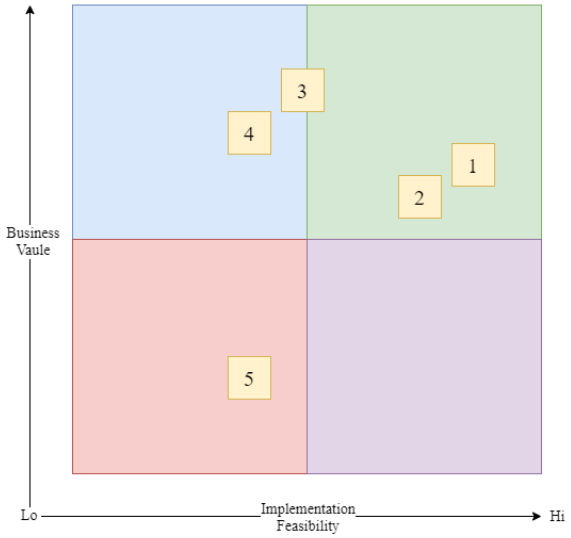


Figure 3: Feasibility Matrix

From here, everyone voted on their favorite use cases taking into account their business value, feasibility estimates and sometimes also personal interests. We also agreed to weight implementation feasibility higher than business value, due to time constraints. Having the best implementation feasibility, and most votes, we decided that the best choice for our analysis was the first use case:

"Where do Verizon's most loyal customers come from?"

## 2.3 Design Thinking

Design thinking is an iterative process that applies *human-centered techniques to solve problems in a creative and innovative way.*[6]

The five-phase model proposed by the Hasso-Plattner Institute of Design at Stanford consists on the following steps: (1) Empathize, (2) Define, (3) Ideate, (4) Prototype, and (5) Test.

The goal is to understand our client's customers' needs and problems, and create solutions based on the insights we have of them. Consequently, improving the quality of the services and the satisfaction of the customers.

Hence, we also focused on design thinking as a complementary problem strategy when completing the business objectives.

## 2.4 Relevant Examples

There are plenty of data-driven decisions around Telecommunication Companies.

To achieve faster and overall better data-driven decisions, multiple companies apply predictive analytics techniques that focus on the data of client preferences and needs. *More than half of the telecommunications companies on the Forbes 500, Forbes International 500, S&P 500, S&P Global 1200, and S&P Europe 350 have also used SPSS software.*[7]

In other words, to better understand the clients, one can use historical data to make forecasts and gain insight. With a high amount and quality of data, the accuracy of the predictive models is usually better.

As a means of motivation, we are going to highlight a couple of examples that proved to be useful in the telco[8] sector that somewhat relates to our business objective:

### 2.4.1 Customer Segmentation

The key to success for telecommunication corporations is to segment their market and target each segment through tailoring the content to them. This rule is relevant to multiple areas of business.[9]

There are four segmentation schemes of prime importance:[10]

1. Client value segmentation: The goal is to calculate customers' value and describe their profiles to monitor needs and preferences.
2. Client behavior segmentation: This is the process of dividing the customers into segments depending on their behavior patterns.
3. Client lifecycle segmentation: It is the segmentation of a customer considering a snapshot of the current life stage of a customer.
4. Client migration segmentation: The idea is to study customer migration patterns and learn when and why a customer ends up in a different segment.

Advanced targeting grants predicting needs, preferences, and customers' reaction to the telecommunication services and products on supply. It allows increased business planning and targeting.

[6] https://www.interaction-design.org/literature/article/what-is-design-thinking-and-why-is-it-so-popular
[7] https://www.spssanalyticspartner.com/industry-solutions/telecommunications/
[8] Used to refer to a Telecommunication Company.
[9] https://activewizards.com/blog/top-10-data-science-use-cases-in-telecom/
[10] https://tasil.com/insights/segmentation-advantages-in-telecom/

### 2.4.2 Customer Churn Prevention

Acquiring a new customer can sometimes be harder than keeping the customers engaged, which still requires a lot of effort.[11]

However, the accurate diagnosis of customers' behavior can alert what customers are at risk of churning. *Churn prediction modeling techniques attempt to understand the precise customer behaviors and attributes which signal the risk and timing of customer churn.*[12]

This allows immediate acting on the satisfaction-related problems and effectively prevents a customer from churning.

### 2.4.3 Lifetime Value Prediction

Customers tend to go looking for better and cheaper services, therefore, the companies need to measure and predict the customer lifetime value (CLV).[9]

Customer lifetime value is a discounted value of all the long-term profits and revenues generated by one client. The CLV model is focused on the purchasing behavior, activity, and services used by the client.

Some solutions collect real-time insights differentiating between profitable, nearly profitable, and unprofitable segments of customers.

# 3 Method

## 3.1 Dataset Description

This dataset contains dozens of features. Some features show information about customer's demographics, such as their age, city, gender, among others. Other features show information regarding their interaction with this company, such as the type of services they use, their monthly charges, the total revenue generated from said customer, and many others.

All this information became available from a Kaggle data set of a fictional company,[13] originally from IBM [14]. This sample data tracks a fictional tel-co company's customer churn based on a variety of possible factors. The data is from the American state California only.

Assumptions:

- Even though the data is fictional, for the sake of this report we pretend that it belongs to Verizon, in order to analyze the business perspective of a business that exists.
- The data set has 7043 entries each referring to a single customer. We will assume that this a representative set of the total number of customers the company has.

By looking into individual features, it is possible to get an idea of what type of customers our client has, to better understand them. It might be useful to know, for example, if the majority of customers are old or young to direct improvement efforts to certain services that the elders or the young are more inclined to use. Some features which provide insight will be looked into in the next section.

[11] https://hbr.org/2014/10/the-value-of-keeping-the-right-customers
[12] https://www.optimove.com/resources/learning-center/customer-churn-prediction-and-prevention
[13] https://www.kaggle.com/ylchang/telco-customer-churn-1113
[14] https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113

## 3.2 Exploratory Data Analysis

We did extensive data exploration to find patterns in data, to discover features that show insight and features who do not, to understand who Verizon's customers are and how they interact with the company. All this was mainly in order to find possible data analysis use cases, and to gauge their feasibility for implementation. To analyze the data, we created Python notebooks that used Pandas[15] and Seaborn[16] for visualization and manipulation.

As said before, looking into customers' individual features will give us insight on who the costumers are and what they are looking for. We did this to find customers' distributions, for a given feature, to possibly find patterns. There's many features in the data-set, which were all looked into, but since the analysis of all of them is very extensive, only a few will be exemplified in this report. First of all, we will focus on their personal information in order to get to know them better:

### Demographics

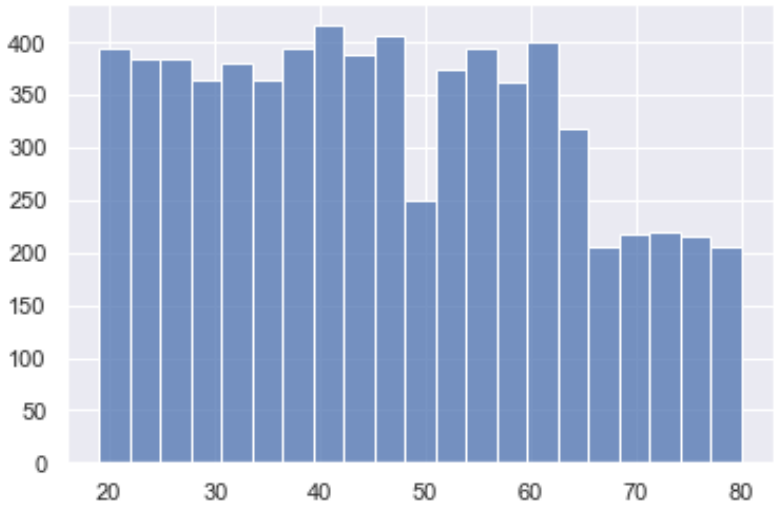


Figure 4: Customer's ages

[15]https://pandas.pydata.org/
[16]https://seaborn.pydata.org/

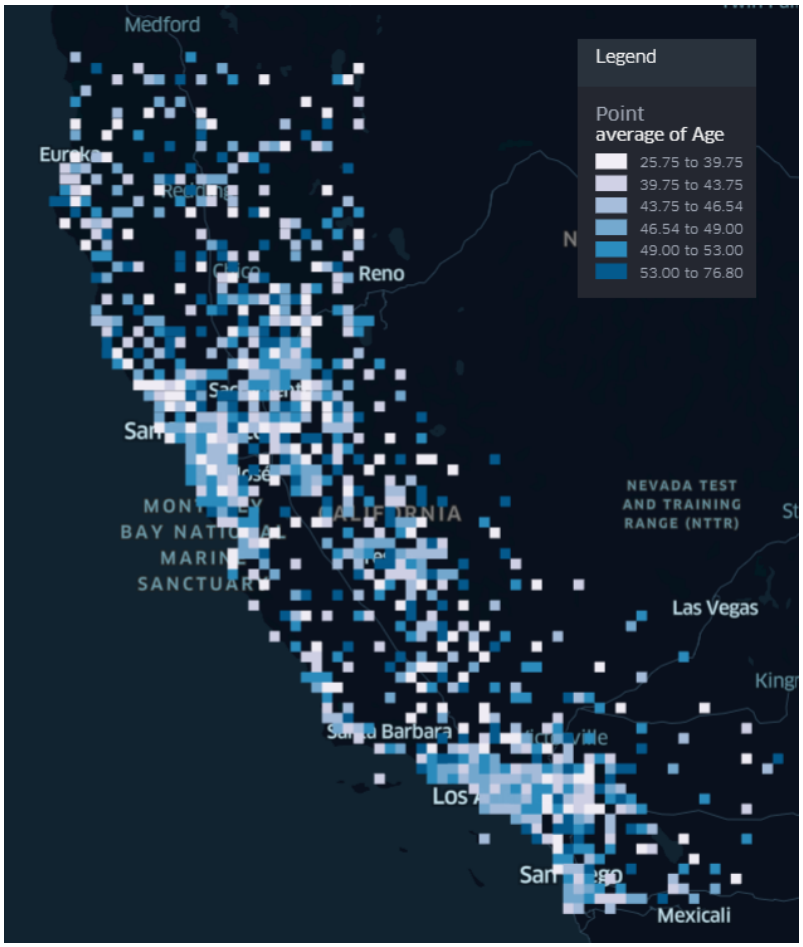


Figure 5: Customer's ages on the map

From Figure 4, we can conclude that the age distribution of customers from age 20 to around age 60 is more or less evenly distributed. The same can be said about ages around 65 to 80. It's noticeable that there's a significant drop of customers around the age of 50. That's an odd occurrence.

Figure 5 shows the customer's ages spread out on the California geography. Each square represents all customers in the 13km x 13km squared area they lay upon, whose color is calculated based on the age average of mentioned customers. This allows us to see that customer's ages do not depend on their location. Big cities and small cities have more or less the same age distribution.

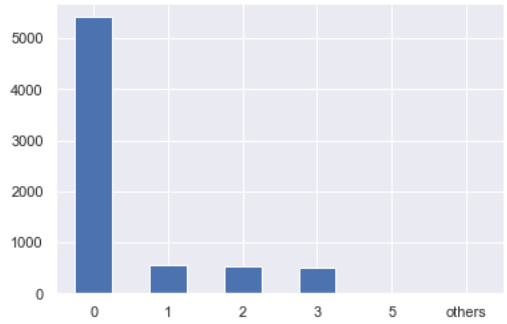


Figure 6: Customer's number of dependents

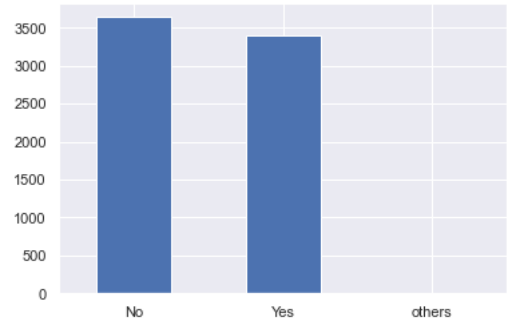


Figure 7: Is the customer married?

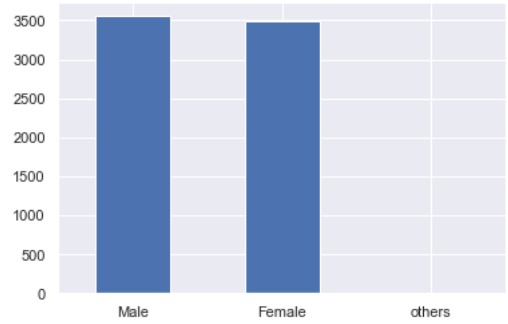


Figure 8: Customer's gender

Figure 6's number of dependents refers to the amount of people (children, parents, grandparents, etc) that live with a customer. We can conclude that most of the customers live alone.

Figure 7 shows there is an almost equal amount of married and unmarried customers.

According to Figure 8, the same can be said regarding gender. Since the numbers are so similar, we can deduce that customers of a specific gender do not show preference for our company.

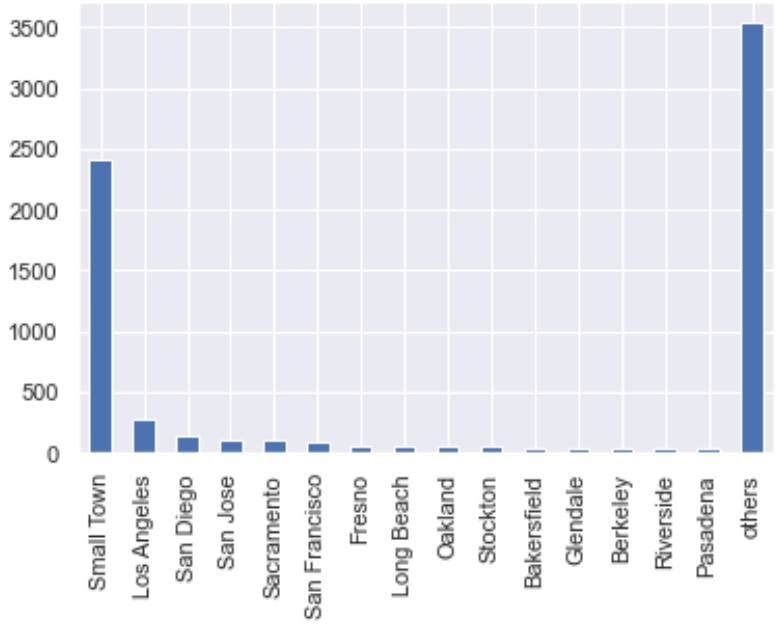


Figure 9: Customer's cities

From Figure 9, it is interesting to see that most of the company's customers come from small towns (which means there are less than 5000 people with the same zip code as the customer).

Looking into customer's locations more precisely, we can see their general location in the form of coordinates. To preserve the anonymity of the customers, each coordinate location is placed in the center of the area associated with their zip code:


Figure 10: Customer's Latitude and Longitude

However, this is not very useful for determining the geographical clusters of loyal customers as all we can see are dots on a map. It only tells us that customers seem to be more concentrated near large cities.

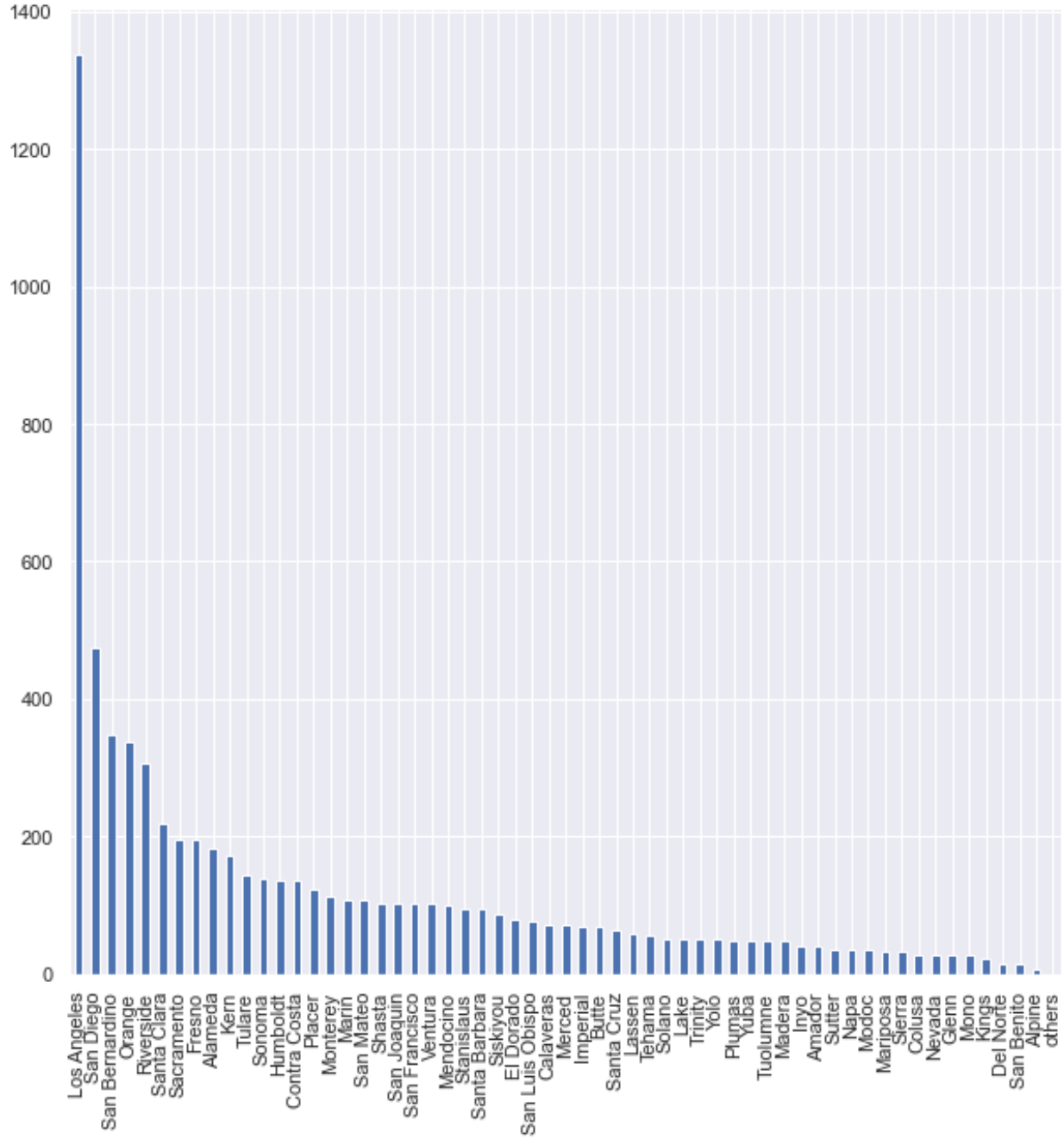


Figure 11: Customers by county

Since there are 1127 cities in California, it is hard to visualize where customers come from if they're grouped by city or zip code. As such, Figure 11 shows them grouped by county[17]. Even more reasoning for this grouping/aggregation is given in Section 3.3. There are 58 counties in total, as shown in the chart. The biggest percentage of customers comes from Los Angeles, whereas the second biggest takes a big leap down, to San Diego. From Tulare onward, the percentage of customers per county slowly decreases, until it reaches almost 0 at Alpine.

[17]The counties are a geopolitical way to divide the territory. California has 58 of them.

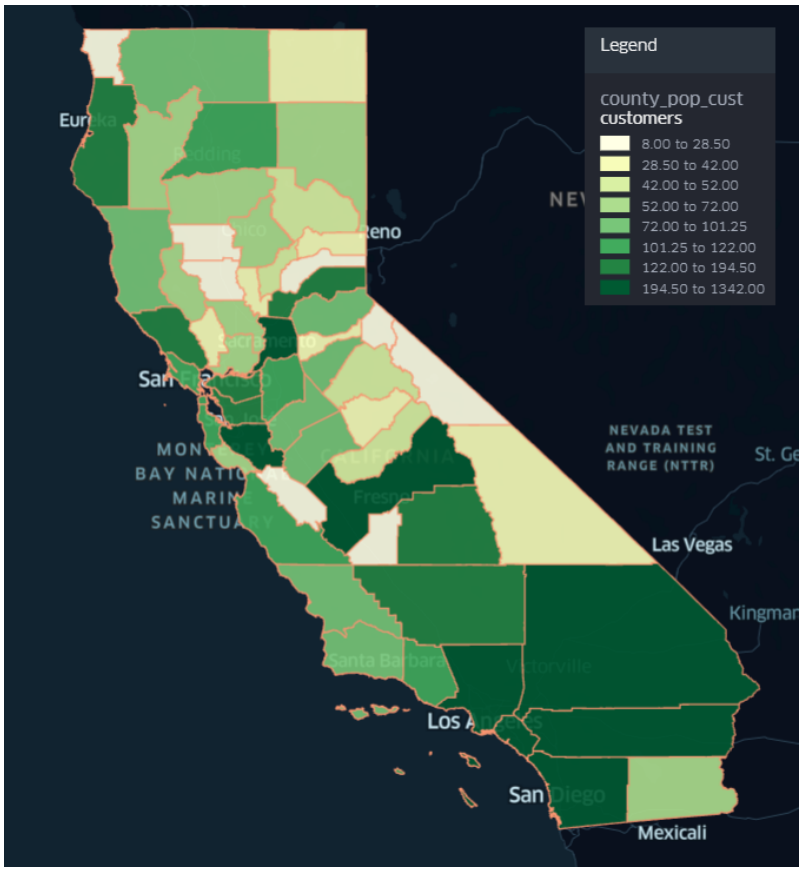


Figure 12: Customers per county

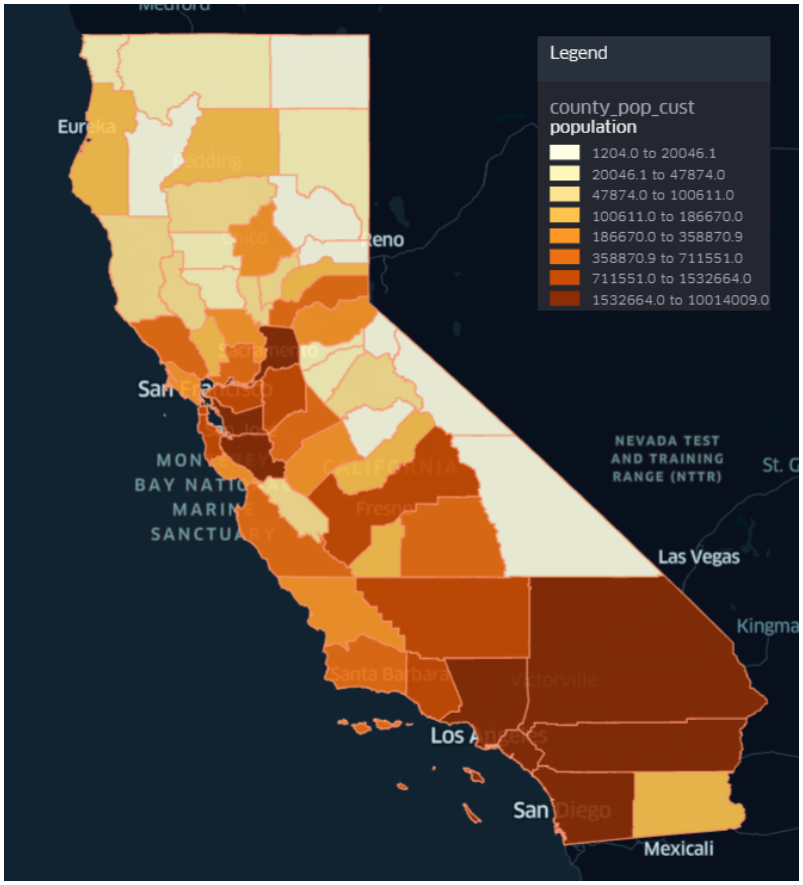


Figure 13: Population per county

The Figure 12 map lets us visualize the customers spread across all counties. The Figure 13 map shows each county's total population. From looking at both maps, it's visible that the differences in color coding between the counties are reflected in both maps. We interpret this as correlation between population size and customer number.

### Customer-Company Features

Now that we have an idea of who our customers are, let's look at how they interact with our company.

Knowing what types of services customers are using will give us a better understanding of what they are looking for.

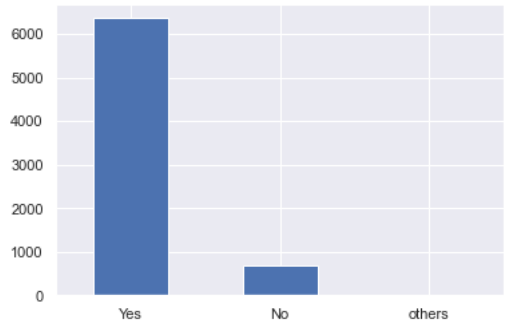


Figure 14: Did the customer subscribe to home phone service?

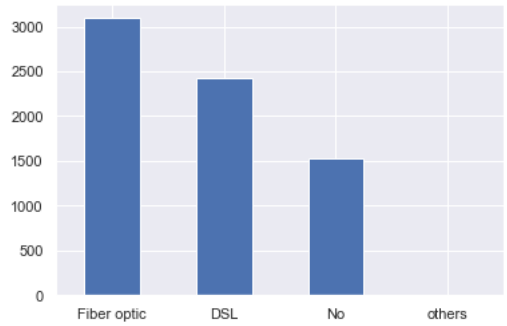


Figure 15: Customer's internet service

Figure 14 shows that the majority of customers use Verizon's home phone services.

In relation to Figure 14, the preference for any type of internet service as shown in Figure 15 is weaker than the preference for phone service.

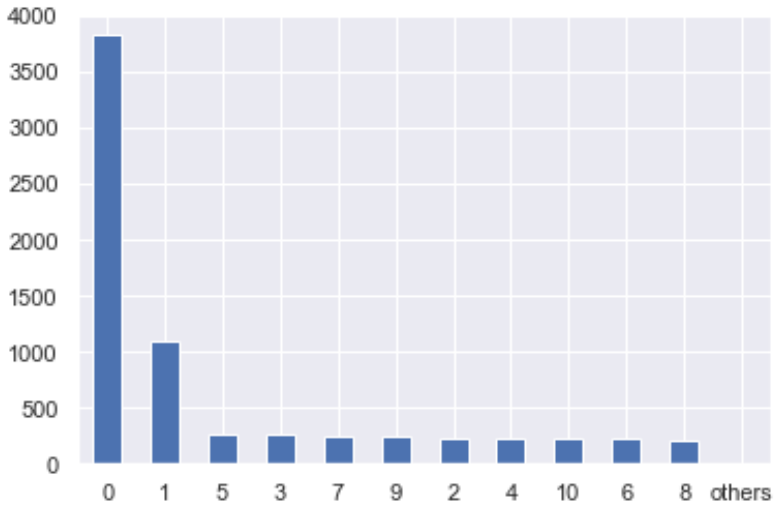


Figure 16: Customer's number of referrals

Number of referrals is the amount of times a customer has referred our company to someone, either to a friend or to a family member.

As shown in Figure 16, the majority of customers have not made any referrals. Even so, a significant amount has made at least 1 referral. It's rather interesting the fact that the 3rd highest number

of customers in this chart are the ones who made 5 referrals, and also the fact that the amount of customers who have made between 2 and 10 referrals are evenly distributed.

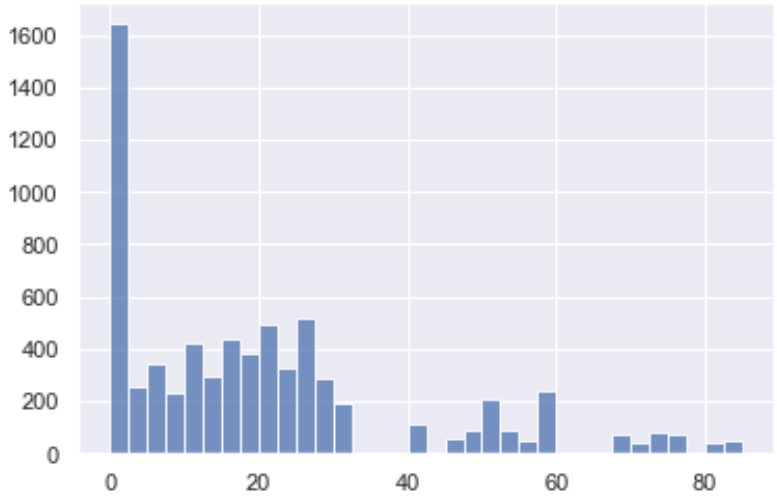


Figure 17: Customer's average monthly GB download

From Figure 17 we can see that a lot of customers barely make any downloads. Another big portion downloads quite a bit, from around 2 GB to around 30 GB. Only a smaller portion downloads much more than everyone else.

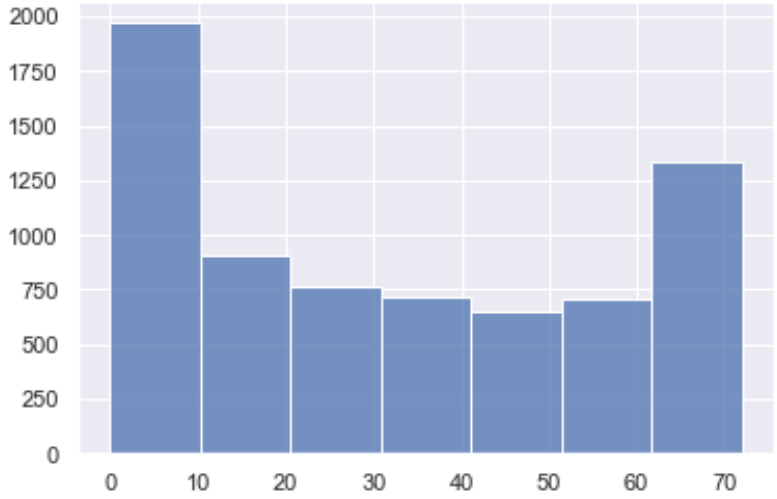


Figure 18: Customer's tenure months

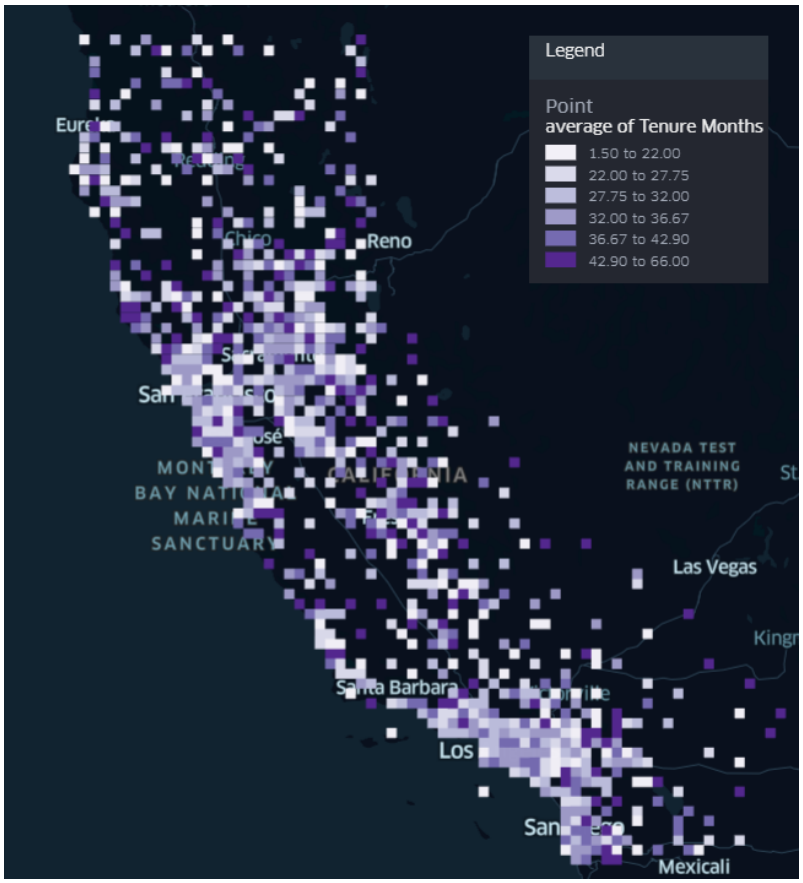


Figure 19: Customer's tenure months on the map

Figure 18 indicates the total amount of months that customers have been with the company. Many customers have been with our client for a short period of time, but most of them have stayed for over a year.

In Figure 19 the same feature can be seen distributed on the map. An interesting pattern that can be seen is that in the Los Angeles area there is a lower percentage of customers who have stayed in the company for a very long time (42 months or more), while the same is not true for other big city areas such as San Francisco.

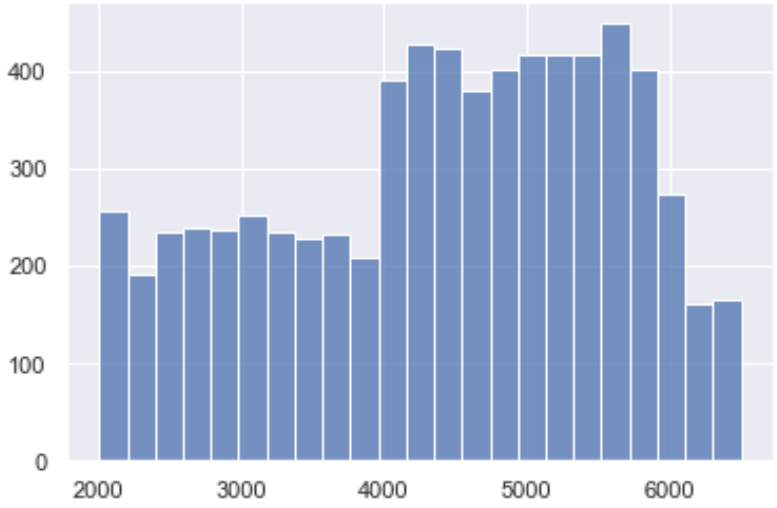


Figure 20: Customer lifetime value

Figure 20 shows a predicted value, customer lifetime value, which is calculated using corporate formulas and existing data. The higher the value, the more valuable the customer. This feature is already provided in the data set. There is a big percentage of customers with a value between 4000 and 6000, so the majority of customers have a high value.

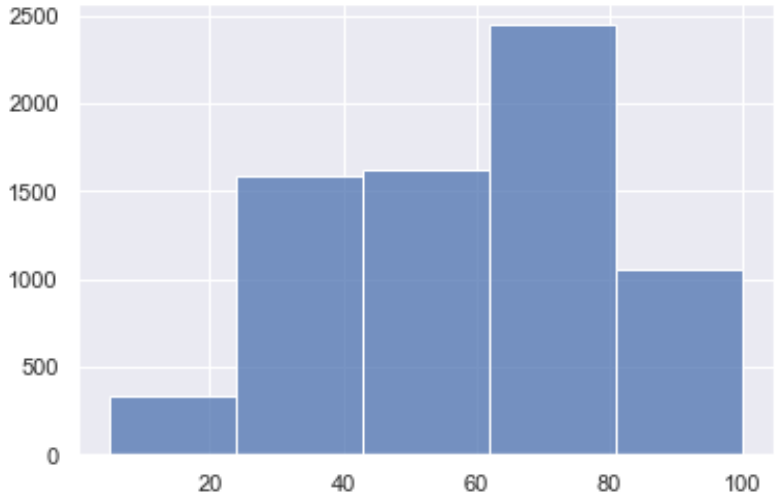


Figure 21: Customer's churn score

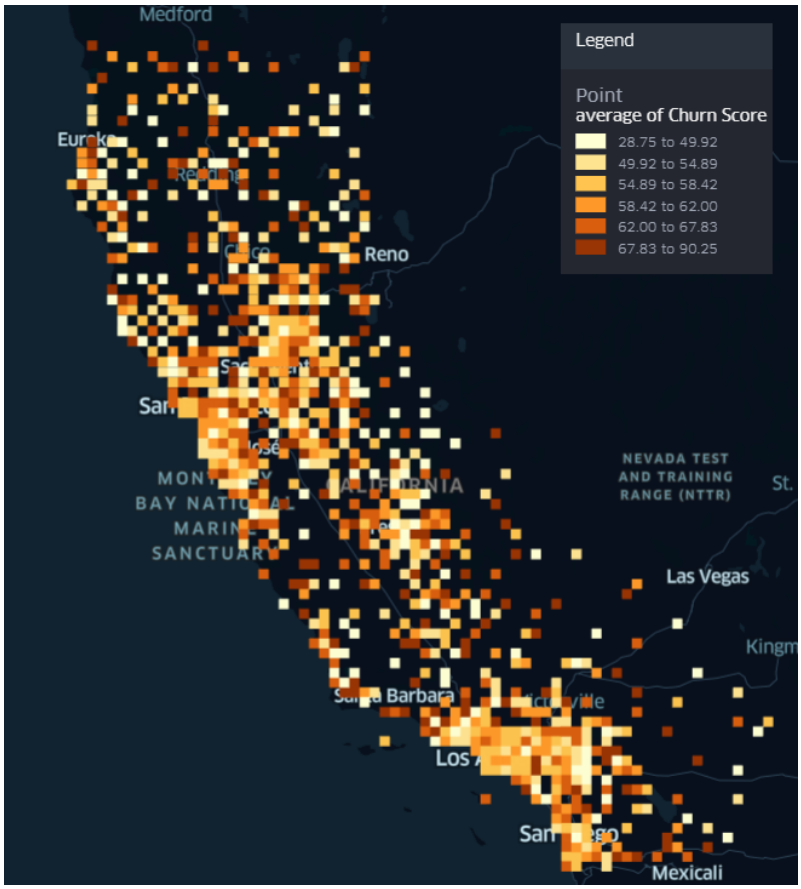


Figure 22: Customer's churn score on map

The churn score, expressed as a percentage, refers to how likely a customer is to churn. The

calculation of such value is done using a predictive tool called IBM SPSS Modeler[18]. The model incorporates multiple factors known to cause churn. This feature is already provided in the data set. Figure 21 shows that the majority of customers has a high churn score. Therefore, most of the customers are at a high risk of leaving our client's company.

The same feature on the map is show on Figure 22. There doesn't seem to be patterns of high or low churn score.

## 3.3 Methods and Tools

In order to derive insights on our use case; "where do the most loyal customers come from?", we first chose *two* angles to the problem through two different definitions of loyalty. Doing this is in line with the description of the CRISP-DM "modeling" phase, which motivates testing multiple models in order to determine what works best (Wirth and Hipp, 2000). Firstly, "most loyal" can be defined as every customer with tenure duration above 30 months, which is roughly the median value of the feature Tenure Months. The median is 29, so it seemed logical to define loyal customers as having a longer tenure than the median. However, any value could be used here. This group is named loyal, and we refer to this approach of loyalty definition long tenure. Conversely, we name the group with the remaining customers non-loyal, just for easier reference. Secondly, the other approach for defining loyalty is through who churned and who did not. This approach is referred to as stayed. Here, those who did not churn *and* have been with the company for more than 3 months are considered loyal. The reason for this 3-month lower boundary is that the dataset specifies whether a customer "joined", "stayed" or "churned" during the last quarter. This means all who joined have a tenure length of max 3 months. Further, examining the different reasons for churn (Figure 1) we decided that those who churned because force majeur factors or important life decisions like the ones in the"deceased" and "moved" categories, should not be considered as part of the non-loyal group, as it would either be impossible or very unreasonable for them to stay with the company. We assume that those who moved relocated to somewhere outside Verizon's service area. Therefore these people are still considered as loyal customers *if* they meet the minimum requirement of staying with the company for more than 3 months. One could argue for other ways of defining loyalty (including "repurchase ratio", which is closely related to tenure months)[19], but the ones adopted here are the most meaningful to us, given our data.

For each of the foundational angles for defining loyalty (long tenure and stayed) they are analyzed by calculating the ratio of loyal customers per county. The reason for aggregating the customers by county (and not by e.g. city) is that it becomes easier to communicate and a more effective visualization can be made, as there are far less counties than cities. The number of data elements in a graph is viewed as affecting the cognitive load (load imposed on human memory) of the receiving audience (Paas and Van Merriënboer, 1994 in Huang et al., 2009). Thus, by aggregating by county only, the cognitive load lessens. The online map visualization tool *Kepler.gl* is used for all our map-plots.[20] Counties with higher loyalty ratios we assume to have a greater chance of bringing in loyal customers to Verizon, in general, without knowing *why* one county has a different ratio than another county. Additionally, it makes more sense statistically to analyze larger customer clusters (like aggregating by county) to avoid statistical fallacy. For example if a small city with only 10 customers has 9 loyal customers, it would give 0.9 in loyalty ratio. On the other hand, by looking only at a county-level might hide information about cities that for some reason are more disposed to give loyal customers. These would disappear if the rest of the cities in a county drags the county-ratio down. This is a clear limitation of our work. We point out that there are roughly 6 customers per city and 121 per county across the Californian dataset, as a high-level average (more on this later in Section 6).

[18] https://www.ibm.com/products/spss-modeler
[19] https://www.capillarytech.com/know-how/customer-loyalty-metrics/
[20] https://kepler.gl

## 3.4 Data Preprocessing

This subsection elaborates on the data preprocessing we did before performing our analyses. First it presents the preprocessing steps, and then how we cleaned the data. The reason for this order is that some data quality errors did not clearly show them selves during our initial exploratory data analysis, but rather during the preprocessing. Therefore the cleaning directly relates to preprocessing steps, and it makes more logical sense to present it after preprocessing.

### 3.4.1 Preprocessing steps

Before starting the main analysis of calculating the loyalty ratios, some further preprocessing first needed to be done.

For each branch of loyalty definition (long tenure vs stayed), the respective loyal and non-loyal groups are extracted. These sets were further filtered, as specified in Section 3.3. Each customer is aggregated into their belonging county, by using their lat/long feature (position coordinates). To help us do this we used geo-spacial geometries of all the counties in California from the GADM[21] dataset. The package *geopandas*[22] is used to check if a customer's position is within the boundaries of one specific Californian county geometry. Counties with less than a 100 customer-entries were skipped from the analysis, to facilitate more representative results. This resulted in 36 counties being skipped, and 22 counties left for analysis. Next in the pre-processing, the count of loyal and active customers (existing customers today) per county was prepared. The count of loyal customers are used for calculating the ratios and the active ones are used for finding the "market share", which is used to calculate the "profitability score" (further explained in Section 4).

### 3.4.2 Data verification and cleaning

The original dataset was divided into multiple Excel files (e.g. *demographics*, *location*, *customer status*. Early in the project we joined them together into one big table, to make it easier to see features next to each other. Later, when grouping the customers by Stayed, we used the Tenure Months feature in the dataset. We discovered that for some customers, Tenure Months and Tenure in Months were not the same number. These two features originated from different Excel files. Using other features like monthly charges and total charges we were able to discover which numbers were the correct ones.

Another issue was that the coordinates were placed in the center of the zip code area. For several areas close to water the coordinates were therefore in the water as shown in Figure 23. As GADM geometries only cover land areas, these customers were removed from our dataset when aggregating over counties. We fixed this by reassigning the coordinates to other zip codes in the same counties.

These issues aside, we found some other minor inconsistencies such as some churn reasons not being assigned the correct churn categories, but none that would affect our analyses.

[21] https://gadm.org/
[22] https://geopandas.org/en/stable/


Figure 23: The point on the map is four customers in the same zip code with coordinate location outside of GADM county geometries

# 4 Analysis

This section will go in depth on the analyses of loyal Verizon customers in the counties of California. The analyses are somewhat similar, differing by their filtering of loyal customers. The section is presented in a logical step by step fashion where each subsection will explain the step in general before going into specifics of the two analyses.

We will use the same definitions of Stayed and Long Tenure as in Section 3.3 when referring to the different analyses and their respective criteria for loyalty.

Please note that in an attempt to avoid severe statistical fallacy, we did not include counties with less than a 100 customers in our analyses. 22 out of 58 counties fit this description and all the plots and figures in this section will only include those 22.

## 4.1 Ratios of Loyal customers per county

After successfully grouping customers by county, we could use Pandas to calculate several different metrics for each county. This included the numbers of customers, currently active customers and loyal customers by both loyalty metrics. Figure 24 and Figure 25 display the top counties based on number of loyal customers.

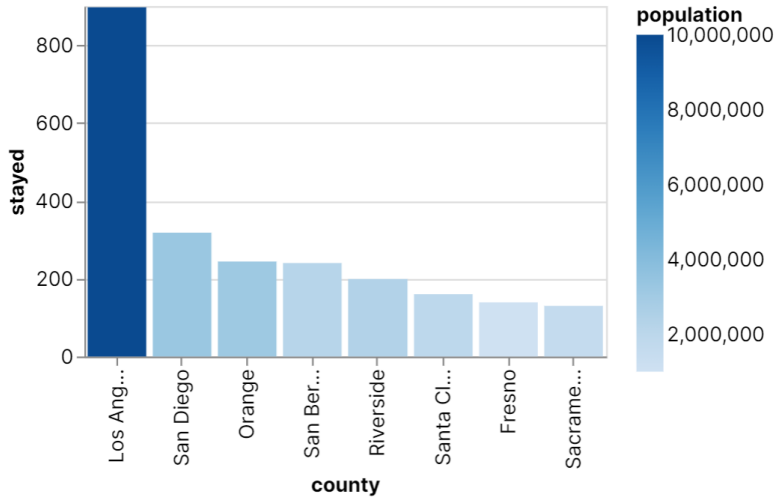


Figure 24: The eight counties with the most Stayed customers

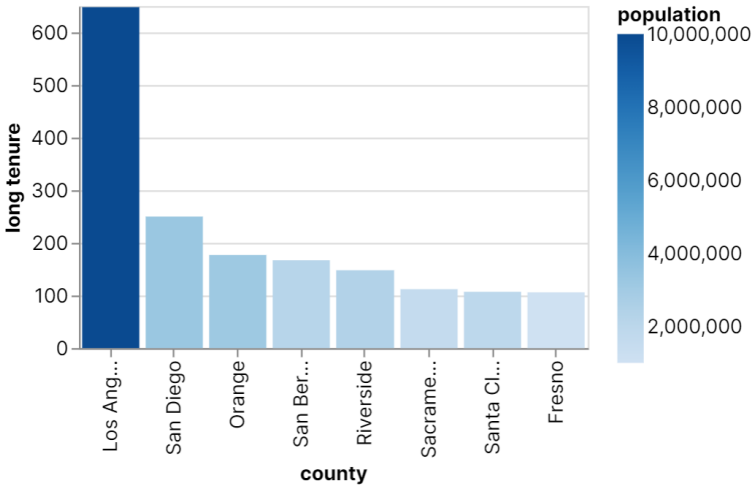


Figure 25: The eight counties with the most Long Tenure customers

We see immediately that Los Angeles would dominate both loyalty categories if we were to look at sheer numbers alone with 897 Stayed customers and 648 Long Tenure customers. All eight counties are in the top ten in population size (Figure 12), and we see that the plots barely vary. It is to be expected that areas with larger populations will have more customers, and likely more loyal customers as well. However this does not mean that Los Angeles automatically is the most effective county to push through the next large scale marketing campaign if Verizon's goal is to increase the amount of loyal customers. Therefore we also looked at the ratios of loyal customers.

### 4.1.1 Ratios: Stayed

The ranking of the loyalty ratio is a much tighter race than the count, as displayed in Figure 26. As we can see, only half of the counties in Figure 24 also have some of the highest ratios. Figure 27 tells us that most of the counties have a ratio somewhere between 65% and 70%. Santa Clara has the highest ratio of 73%.

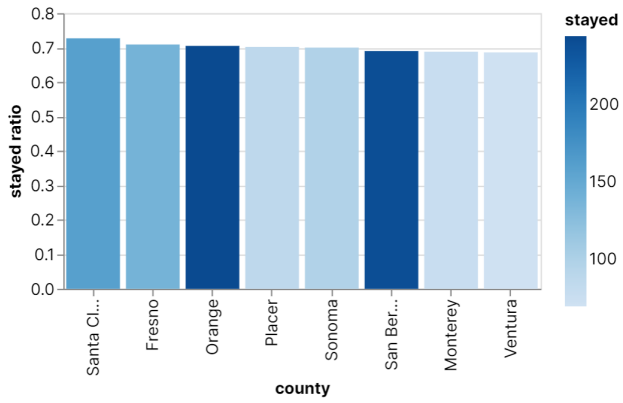


Figure 26: The eight counties with the highest Stayed ratio. Santa Clara is on top

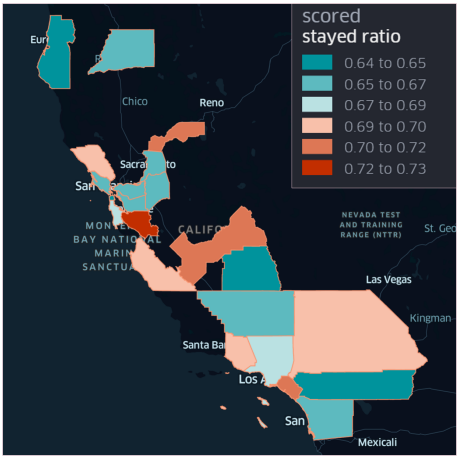


Figure 27: A map showing all 22 counties with a color mapping based on their ratio of Stayed customers. The Dark red county is Santa Clara

### 4.1.2 Ratios: Long Tenure

Here as well the ratios are a lot more evenly matched than the counts. (See Figure 28). The difference between count and ratio is not as substantial for Long Tenure. Six out of eight of the counties with the most loyal customers are among the top eight when looking at the ratio. Figure 29 shows that most of the counties in the analysis are clustered around a 50% ratio, which is quite interesting. The Long Tenure metric criteria is as previously discussed based on the median of tenure months among all customers. This plot reveals that the distribution of customers by tenure months across individual counties is mostly quite similar to the distribution for the entire dataset. The highest ratio of 57% is found in Sacramento.

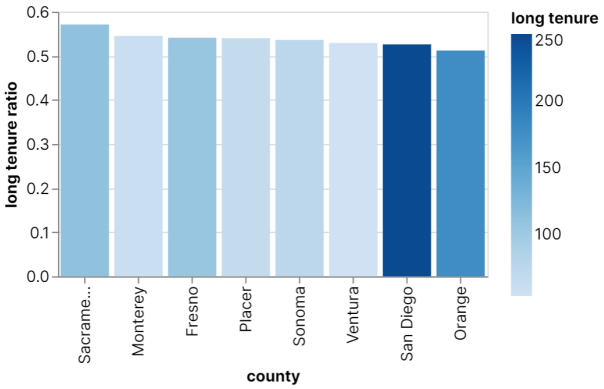


Figure 28: The eight counties with the highest Long Tenure ratio

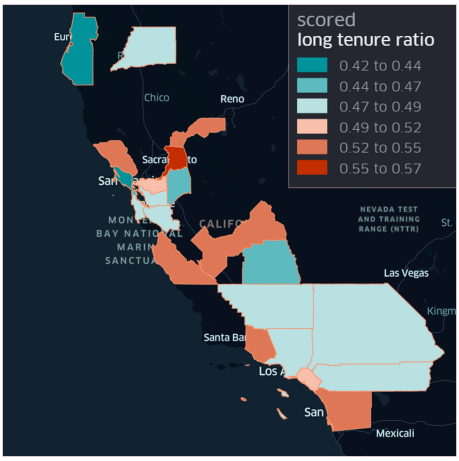


Figure 29: A map showing all 22 counties with a color mapping based on their ratio of Long Tenure customers. The Dark red county is Sacramento

Note that even if Los Angeles dominates the number of loyal customers, it is not in the top eight of either ratio ranking.

## 4.2 Profitability score of marketing campaigns

The results from our analysis related to the use case *"Where do Verizon's most loyal customers come from?"* in the form of ratios might have their intended effect of reducing churn rate by increasing the amount of loyal customers if Verizon decides to launch marketing campaigns in the top rated counties. However, there are other demographic factors than just the loyalty ratio that influence the efficiency of a campaign in regard to maximizing loyal customers. Perhaps the most important one is the company's growth potential in the county. If the highest ratio of loyal customers is in an area where Verizon has 90% the market, it could bring in more customers to launch the campaign in a county with a lower ratio, but where Verizon has a smaller presence. While this is not a problem if we assume the total customer count is around 7000 customers for all of California (as the market (as the market shares are insignificantly low), it could be important to think about for future decisions if the company grows.

### 4.2.1 Devising a Score

As outlined in the previous section, basing our recommendations on the number of loyal customers alone would have a severe weakness as these are heavily influenced by the population size of the counties. Using only the ratio however would ignore the benefit of having a large potential growth of loyal customers. Therefore we devised a scoring metric reflecting the profitability of marketing campaigns in counties, incorporating both the loyalty ratio and the growth potential in the form of current market share. The final profitability score is defined by Equation 1.

$$Profitability\ score = Ratio * (1 - Market\ share) * 100 \tag{1}$$

The multiplication by 100 is only added so that the numbers will end up in the 0-100 range. This helps to make a clear distinction from the ratios which are fractions.

Because our dataset only has about 7000 entries and the population size of California is over 39 million, our market share numbers became very insignificant. For the growth potential to actually matter in the equation we boosted the market share (by 1000) so that it would be in the whole percent range, i.e. closer to one percent, for most counties. The market share was therefore calculated using Equation 2.

$$Market\ share = \frac{Active\ customers}{County\ population} * 1000 \tag{2}$$

It should be mentioned that Humboldt county gets a market share of 76% because of the boosting, which might be unrealistically high. It is the county with the lowest number of inhabitants among the 22 in the analyses and it does have a significantly higher market share than the rest of them. As the main point of this boosting is to make market share significant and it was the only county with this problem, we deem the lower score as a result of boosting justified.

### 4.2.2 Score results: Stayed

We see from Figure 30 that while Santa Clara still takes the lead with a score of 66.5, the market share really impacts the ranking. Again only half of the counties in the top eight ratios are also in the top eight in profitability. It is also clear that the lead for Santa Clara is larger now, indicating that it is a county with a good ratio and a large growth potential. The map in Figure 31 shows that the overall difference is larger for the profitability than for the Stayed ratio even when not considering Humboldt which significantly pulls down the lowest category with a score of 15.2. The bulk of the counties score very close to 60.

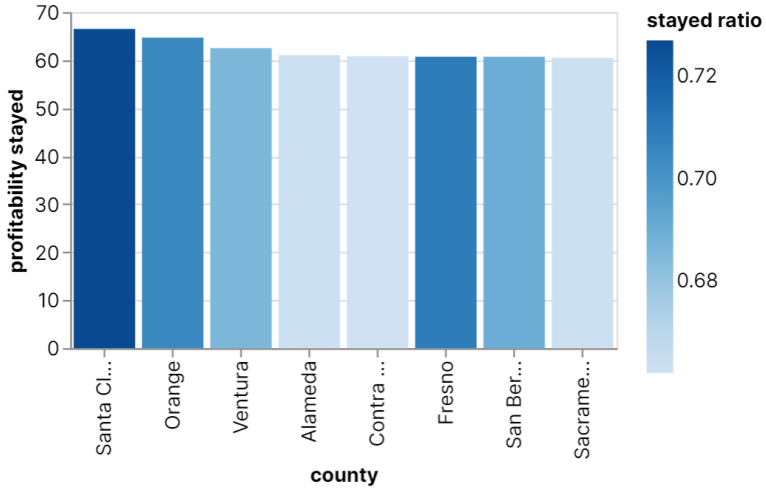


Figure 30: The eight counties with the highest Stayed profitability

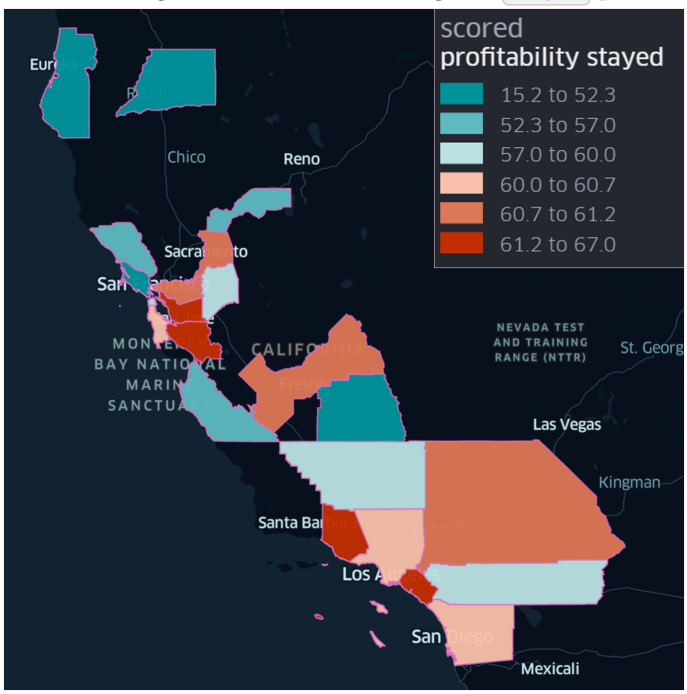


Figure 31: A map showing all 22 counties with a color mapping based on their profitability calculated with the Stayed ratio

### 4.2.3 Score results: Long Tenure

For Long Tenure as well, the leader of the ratio also has the highest profitability, signifying a favorable growth potential. The span of the profitability score is marginally smaller than the span for the ratio when not considering Humboldt. Sacramento leads the ranking with a profitability of 52.1.

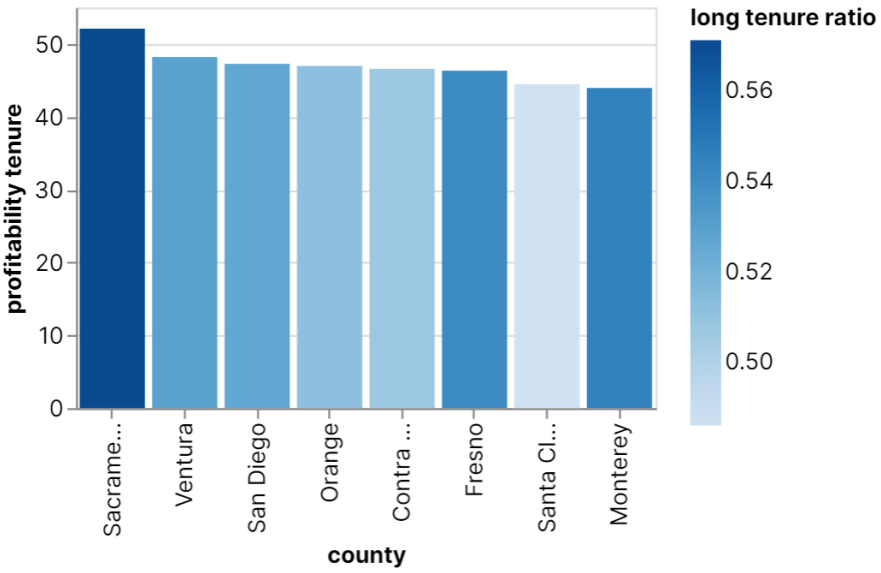


Figure 32: The eight counties with the highest Long Tenure profitability

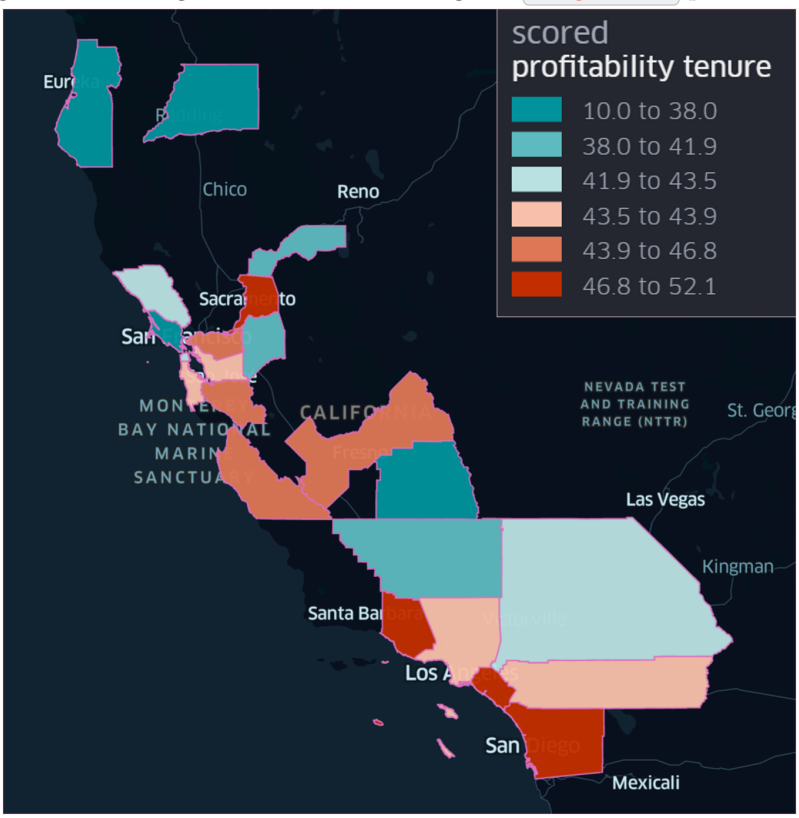


Figure 33: A map showing all 22 counties with a color mapping based on their profitability calculated with the Long Tenure ratio

# 5 Interpretations and Recommendations

After doing the data analysis, through digging into the two definitions of loyalty, and looking on the profitability scores, we came up with two main recommendations: focused marketing and analyzing the loyal counties. These are further explained below.

## 5.1 Focused Marketing

We recommend our client to focus their marketing in the counties with the highest profitability score in order to maximize the number of loyal customers there, since it would increase the final revenue the most. Thus, we strongly encourage our client to launch marketing campaigns in Santa Clara and Sacramento, which are the counties with the highest profitability for Stayed and Long Tenure respectively. After these two, it would be a good idea to consider launching marketing campaigns in Orange, Ventura, Contra Costa, San Diego, Fresno and Alameda as these are the runner-up counties that have the highest average profitability score (across Stayed and Long Tenure). See Figure 34.

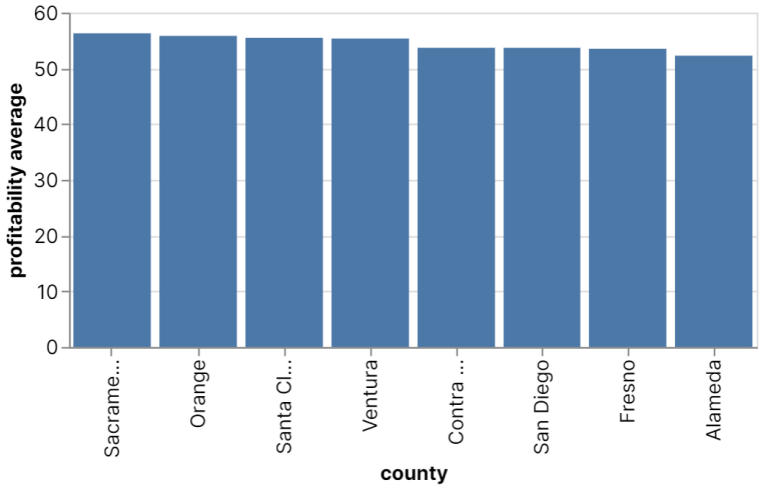


Figure 34: The eight counties with the highest profitability average

In line with step 5 of the CRISP-DM model, we should chose the approach we think will serve our business goal better. Our opinion is that the better metric to focus on is the long tenure metric. While some of the customers considered loyal in this category have churned, 30 months is a long time, and a 30 month subscription generates a lot of revenue for the company. Therefore focusing their marketing in the counties where the Long Tenure metric is high Should generate more loyal customers who will stay for a long time according to the ratio and score. These are the most valuable customers for the company in terms of revenue, which is the bottom line goal of increasing the number of loyal customers and reducing churn rate.

## 5.2 Analyze Loyal Counties

A further and more detailed analysis on the counties with higher loyalty ratios would give our client a better perspective and understanding on which factors inspire and contribute to a high number of loyal customers there, letting them know what makes the customers stay longer periods and/or stay in the company. Gaining insights on such factors allows Verizon to take action and try to improve the loyalty ratio in those counties with lower loyalty ratios, for instance through tailoring a new marketing strategy for the counties with low loyalty ratios. In other words, a deep knowledge on what makes a county have more loyal customers can be a clue to increase the loyalty ratio in other counties.

## 5.3 Implementation Plan

The table below outlines our recommended implementation plan for Verizon, for the upcoming months.

| Recommendation | Stakeholder | Time-frame |
|---|---|---|
| Launch new marketing campaigns towards all of the recommended counties | Marketing | Jan 2022 - March 2022 |
| Analyze customer profiles within all the recommended counties to find commonalities | Analytics | Jan 2022 - March 2022 |
| Launch targeted offers toward non-customers that match the profiles found above (for counties with high potential) | Marketing | April 2022 - May 2022 |

Table 3

# 6 Limitations

In this section we will discuss some limitations that exists in our work.

When viewing loyalty as having been with the company for more than 30 months, one customer who has currently not been a subscriber for that long might actually end up staying with the company beyond this limit when the time has passed. As such there might exist many customers that *are* "loyal" customers, but haven't had the chance to prove themselves yet, and therefore they are not included in our loyal set. On the other hand, as long as the company has had their offerings on the market for a sufficient time, one can also assume that the relative differences in loyalty inter-county will still be reflected in both the loyal and non-loyal groups. This is of course dependent on that there *are* latent factors associated with different counties that affect customer loyalty, which our analysis indeed suggests. It is also possible that the differences are based on chance (more on this below). Out of limiting our scope, we do not go into what those factors might be connected to, but this would be interesting to investigate for future work.

Moreover, choosing 30 months as the boundary for a customer to be defined as loyal (in the Long Tenure approach) is basically arbitrary (but it is roughly the median of this data feature). Any number could be chosen, and Verizon might want to determine another value which might be more meaningful for them.

We don't see any significant weaknesses connected to the Stayed definition of loyalty itself. Adding the criteria that customers have to have been with the company for a short period is meaningful (yet 3 months as this boundary is chosen, arbitrarily, but based on convenience, given our dataset). It is however possible that it would be more meaningful for Verizon to choose an even higher value for this lower limit.

Further, the assumption that a *new* customer is going to have a higher probability of being loyal for a county with higher loyalty ratio—may not hold. This is related to the assumption that the differences in loyalty are majorly explained by the existence of some latent factors in the population of a county. However if this is not the case, then betting on getting more loyal customers based on this ratio is equally weak of a strategy as using solely the historic development of a stock's price to determine the future development of that stock (that if a stock increased $X\%$ one year, it would increase $X\%$ the next).

One highly significant limitation, as briefly introduced earlier in this section, is that the loyalty ratios observed might be due to chance. This can for instance originate from one county having too few customers reported for us to be confident in their loyalty ratio. As an example, 16 of the 22 counties that were included in the analysis had only between 100 and 200 customers. This is compared to 1342 customers in the county with the highest amount of customers (Los Angeles). For perspective, the second largest county had only 476 customers, and only 6 of the 22 counties had customers above 200. This means most of the ratios were probably not very representative, judged by intuition (though, more representative than if we had not skipped those with less than 100 customers). One possibility here would be to measure the confidence interval (the uncertainty boundaries) of each of our calculated ratios (one for each county), based on the amount of customers in each county. This would give a clear answer to how much we could trust our results. This footnote showcases how this could be done, if we replaced the machine learning

*accuracy* score with our loyalty ratio.[23] These findings mentioned here should have been included in our presentation-video, but was not discovered until the final stages of writing this report. Another possibility related to this would be to integrate the amount of customers in the county into its profitability score. This way the score would be affected by the strength of the evidence.

Furthermore, since we are aggregating by county and doing our ratio analysis on this level, some of the 1515 Californian cities' potentially extreme ratios might go hidden. This can happen if some cities contribution to the county's ratio cancel each other out (some highs are canceled by some lows). Therefore it could be interesting to do the analysis on a city level, even though this would be infeasible to visualize in a pretty way.

To finish, the dataset clearly provides too little data overall, since 36/58 counties were excluded due to having less than 100 customers. For these counties we might miss out on some that could have turned out to be very valuable to Verizon.

# 7 Conclusion

Our data exploration gave us information about who our customers are, what services they bought, why some of them churned, and so on. This analysis allowed us to discuss and come up with the business objective of increasing Verizon's loyal customers. Five use cases were also created with this information in mind, out of which we ended up choosing "Where do Verizon most loyal customers come from?".

To tackle this use case, two definitions of customer loyalty were created, the first being defined as long tenure and the other stayed.

Before doing calculations, we did more pre-processing such as discarding the counties with less than 100 customers. Only then did we calculate the ratio of loyal and active customers per county. These calculations were used to find the "market share" which was used to calculate the "profitability score". This score reflects the profitability of marketing campaigns in counties (a higher score means more profitable). It consists of a modified market share fraction due to the low amount of customers in the data-set and loyalty ratios per county. Based on counties' amount of customers, currently active customers and loyal customers, we obtained the top counties for both definitions of loyalty: Santa Clara and Sacramento.

With the data analysis and profitability score complete, we were able to make two recommendations:

- to focus the marketing on counties with the highest profitability score, as to increase the number of loyal customers,
- and to analyze with more detail the counties with higher loyalty so that Verizon may understand what factors contributes to a high number of loyal customers.

Finally, we recognize that our work has some limitations such as:

- The two loyal customer definitions exclude customers who might be loyal but haven't yet proved themselves, and include customers who might churn after a short period of time.
- We assume that a customer is more likely to be loyal if he comes from a county with a high loyalty ratio. That might not be true.
- Loyalty ratios might originate from chance due to the low amount of customers in some counties. Overall, most counties had a low number of customers, and we are therefore unsure about the strength of our results.
- We decided to filter out counties with less than 100 customers. This makes it possible to miss out on some counties that could have proven themselves very useful to Verizon.

[23] https://machinelearningmastery.com/report-classifier-performance-confidence-intervals/

With this summary, our analysis is complete.

# Bibliography

Hindle, G. A. and Vidgen, R. (2018). Developing a business analytics methodology: A case study in the foodbank sector. *European Journal of Operational Research*, 268(3):836–851.

Huang, W., Eades, P., and Hong, S.-H. (2009). Measuring effectiveness of graph visualizations: A cognitive load perspective. *Information Visualization*, 8(3):139–152.

Paas, F. G. and Van Merriënboer, J. J. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational psychology review*, 6(4):351–371.

Schmarzo, B. (2019). *The Art of Thinking Like A Data Scientist*.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1. Springer-Verlag London, UK.