Insights:

1. **Rare/Out of business makes**: from the looks of the graphs, people who are buying more fringe cars that can’t be found elsewhere or are more rare have a much higher loyalty rate. Some of these cars, for example, are daewoo, eagle, GEO, isuzu, plymouth, oldsmobile, and SAAB.
   1. Made a really great graph for this, I think. I replaced the purchase\_make with the relative popularity of the make across the dataset. Super rare cars have great loyalty and then it tapers off.
   2. Maybe also find avg dist to dealer to support this
2. **Financing and makes**: we find that people who don’t use financing and buy more high end vehicles have a larger loyalty rate => car aficionados who have money.  Some of the examples for this are audi, bentley, hummer, jaguar, pontiac, and porsche. Likewise, these cars may also be harder to find (i.e. pontiac because they aren’t made anymore, jaguar because british and few dealers) or better deals can be achieved at carmax due to it being resale
3. ***Strategy for the two above****: targeted online marketing.  If people know you have these rare cars, high end cars that are hard to find, or high end cars for a better deal, you will likely get more traffic and by extension more loyal customers.*
4. **Car Price:** we see two things in purchase price that can help confirm two theories.  First off, $0-5000 cars have very high loyalty rates, these are probably people who buy cheap and have a high car turnover rate or 2 are young people and are going from buying their first car and later buy another car.  Second, higher priced cars are better more loyal. Once again, i believe this has to do with aficionados, people who like to buy a 2 year old car and trade it in or buy a new one in just a couple years. 1. Heat maps on age and price would be dope 2. Heat map on income and price would be good
5. **Higher income w/ trade in = more loyalty**: we find that higher income people who have trade in have a significantly higher likelihood of being a loyal customer.  This also can add to our afficionado thesis where they’re buying more frequently and trading in so they can have more different kinds of cars.  It would be good to check this agianst 1. Whether the known income was financed 2. What cars they were buying but that could be hard.  Check slide 11 to see what I mean
6. **Trade in on older models (& brand new) are higher**:  old cars have a much higher loyalty rate. This is likely due to the fact that they have a high turnover rate because they die quicker and people eventually need to get a new car.  Also, people who trade in and buy an old car are more likely to be loyal, this leads us to believe that they might repeat this cycle, but data on the subsequent purchase would be needed.  Outlier => people who buy brand new (2014) and trade in had major separation.  Probably people had an older car but were ready to get their shiny new car.
7. ***Strategy for the 2 above****: have a good loyalty program based around trade ins.  Maybe the more you trade in, the better deal you get on your next purchase.  You lose some in profits but the gain you could achieve in repeat customers could be very beneficial.  1. Check trade in over all unique # of subsequent purchases 2. Check trade in with age as well, its likely young people buy a first car and then trade in and get a second*
8. **Distance and strategy - Expand the neighborhood:** we found only two real telling data points with distance.  First is that people within a mile of the store are far more likely to be loyal. It seems likely that the people who live right next to a carmax will think about it much more since they see it all the time and will then buy cars from the branch.  One way to increase the “neighborhood” is be active in the immediate communities around a branch. It’s more than advertising when you help sponsor an event, partake in community service, or present opportunities like this showcase to highschool students around branches.  The more people near by think about you, you could get more loyalty.
9. **Null distance**:  one anomaly we found in the data was people you don’t have a distance for have a 100% single subsequent purchase rate.  This is very bizarre.

Script:

**Slide 1:** Introductions

**Slide 2 (Methodology):** Our strategy when approaching the problem of determining customer loyalty has four steps.  First, we take an initial look at the data for distributions, missing data, and relationship to our response variable of customer loyalty.  After the initial look, we dive into statistical processing and machine learning to see if they can guide our search, validate any initial assumptions, and tell us which features have the greatest relative importance.  After machine learning, we dive back into analysis with the insights taken from machine learning and check out two factor analysis as it relates to loyalty. Finally, we gather our thoughts and findings together into some business insights and strategies to increase loyal customer rates.

**Slide 3 (Initial Look):** We found 6 features with missing data, and handled this in different ways based on the feature (brief summary of how you handled).  We also found that on average 34% of the population would make a subsequent purchase.

**Slide 4 (loyalty definition)**: We defined loyalty as anyone who purchases 1 or more car in the following 5 year period.

**Slide 6 (gender and age single feature):** Males held the majority of purchases and were found to have a slightly higher rate of loyalty while age showed that the 0-20 and 101+ had the most significant difference from the mean but 21 - 61 held the vast majority of purchases.

**Slide 7 (Income and previous purchase single feature):** We found people who make less than 80k buy the majority of vehicles but with increasing income >= 140k we see loyalty rates start to rise.  Previous customers also were more likely to remain loyal, however its worth noting only 22% of the population was a previous customer which is less than subsequent. This could be do to lack of centralized data systems to identify the information

**Slide 8 (vehicle make and price single feature): MAKE PERCENTS LABELED WRONG** When bucketed into major categories of cars, we see little difference but buy looking at purchase price ranges individually, we find $0-5k and $55K + cars have a higher loyalty rate.

**Slide 9 (year and warranty sing feature): YEAR GRAPHS LABELED WRONG** Purchases prior to 2000 are almost negligible in count but 2000 - 2003 have a noticeably higher loyalty rate.  We also find using a warranty increases loyalty rates by 2%.

**Slide 10 (distance to dealer, trade in, and financing single feature):  TRADE IN % LABELED WRONG** When addressing distance to dealer, only those within a mile greatly deviate from the mean with a much higher customer loyalty rate.  People who don’t finance tend to have higher loyalty rates at 37% and those who trade in also have higher loyalty rates at 36%.

**Slide 12: talk about your chi squared test**

**Slide 13:**

**Slide 14:**

**Slide 16 (make demand single feature):** This graph here shows the relationship between car makes and loyalty, however the X-axis is oriented to show demand increasing from left to right. We can see that niche cars seem to have a much higher rate of loyalty among customers.

**Slide 17 (financing, make vs loyalty double feature):** The graph we have here shows car makes split into financing and non financing and it’s relationship to loyalty. The key takeaways here are that high end cars bought without financing have higher loyalty rates. We can infer from this that these are high income people and probably buying rather frequently

**Slide 18 (insight & suggestion):** in the world of online marketing, there are tons of companies whose entire job is getting the right adds to the right people. In this case, advertising that you have these rare, niche, or used high end cars could attract both new customers who didn’t think carmax might have such a vehicle or bring in old customers who we already know are more likely to be loyal and just need a little push to get in the door. We like to call them aficionados because they aim for the rare cars and seem to be loyal because you have them.

**Slide 19 (price, age vs loyalty double feature):** We see above average loyalty rates amongst people ages 0-20 for cars under $45k. We also see that $0-5K cars have higher loyalty across the board. Interestingly enough, we also see 101+ people being exceptionally loyal customers (can’t take it when you’re gone I guess)

**Slide in between these (purchase price, # subsequent purchase vs trade in):** Here we see that people who have many subsequent purchases are both buying high valued cars and trading in cars. This is just another way to see that high end vehicles and trade in programs can be a very lucrative business for carmax.

**Slide something (insight & suggestion on trade in loyalty program):** Two things we saw in our analysis around trade in is first, people who trade in and buy old models (2000-2002) have a much higher likelihood to be a loyal customer. We might assume this to be a repetitive cycle: buy an old car, drive it for a couple years, trade it in for what it is still worth, and buy another old but slightly newer car. The other insight we found with trade ins is there is a positive direct relationship between loyalty and income of people who trade in. These feel like people who like to buy a new used car every couple of years. To service both these groups, a good loyalty program around trading in could be put together. Maybe every so many trade in and purchase you get a little cherry on top to make it all the more worth while to keep coming back. There are many ways you could implement this, that is just one.

**Slide something (insight and suggestion expand the neighborhood):** Customer distance to dealer was rather important according to machine learning and one reason for this could be the drastic difference in loyalty of people within a mile than every other distance. These are people who probably drive by your branch on a daily basis so it is always fresh in their mind and keeps them likely to shop there. One way to expand this “immediate neighbor” radius would to be to get active in local events. Maybe have branches participate in community service, sponsor some HOA event nearby, or present some interesting challenge (like this showcase) to a local highschool. Anything a branch can do to get more personal with people in the community will likely be far more influential then most advertising could be. You want to be in their head, the first place they think of.

Methodology report:

To complete the analysis of a medium sized data set like the one presented to us in this challenge, our team chose to use python as the medium to work with the data.  Python is ideal because it has many very data centric built libraries like pandas, sklearn, matplotlib, seaborn, and more. Jupyter and Atom with Hydrogen were used as UI for running python since they allow you to run code in cells and display visualizations on screen.

When it comes to our methodology for analyzing the actual data, it follows our presentation outline exactly.  We do a preliminary exploratory data analysis to understand the data at a high level.  Next, we utilize statistics and machine learning to derive unintuitive insights.  Next we go back into analysis and check two feature relationships against the response variable. Finally, we take all of the information in and synthesize it to come up with business insights and strategies.

In preliminary analysis, we began by looking for missing data. Missing data can skew results in analysis and machine learning if not addressed so it is the first step. We found missing data in 6 different features and dealt with them on a case by case basis. Post satisfaction survey was missing in approximately 360,000 of the 366,000 records, and after finding no important relationship with loyalty, was omitted for the remainder of the analysis. Customer income and age were filled with the median. It is worth noting that data of missing customers had drastically smaller levels of financing. Customer distance to dealer and customer gender had missing records that were filled by the distribution of data points. It is important to note that all missing distance to dealer had exactly one subsequent purchase. Finally, there were only 3 missing price records so these were dropped as well.

Our initial look into the data included checking the distributions of different features to know what populations control the means. Once we see distribution, we check the features against the response variable to find the rates of loyalty within a given population. Stacked bar charts and area charts were used to make these visualizations. After preliminary analysis, we generated a correlation matrix to understand how features worked were related. We next used machine learning to generate predictions and perform chi squared correlation to find pairwise significance for each feature against the response. This is beneficial because analysis can find trends in data but machine learning can tell us its relative importance. After machine learning, we address two way feature relationships to the response variable. This is done by finding the intersection of two features, calculating loyalty rate between the population within the two features, then mapping to a heat map so it is easily digestible.

Synthesizing the machine learning and straight visualized analysis, we start to generate our insights and suggestions. We try to focus on the important features and those with heavy trends, and once identified, build customer profiles around these indicators of loyal customers. Once a customer profile is built, we can develop strategies around meeting their needs to further service them and retain them as a loyal customer.