Supervised Capstone Presentation

BY BENEDICT LAI

JULY 14, 2019

Introduction to Dataset

Why I chose this data set?

This data set had many categorical variables, which gave me the opportunity to tackle more challenges. Also, I was interested in finding out the type of age group that participates the most when it comes to Black Friday shopping, meeting up with friends, participating in Friendsgiving, etc.

What do I hope to find out?

I want to predict a person's age based on their Thanksgiving habits.

I want to project if younger people enjoy spending time with friends more on Thanksgiving to reunite with high school classmates.

I want to make a prediction if younger people working in retail correlate well to those who do Black Friday shopping and meeting up with friends on Thanksgiving night.

I want to predict if there a trend with younger demo when it comes to income annually.

What does success for this project look like?

What will make this project successful is to predict what age groups are the most popular with each other question and what kind of eating habits does each age group have. Success will look like producing a model that is better at predicting than random guessing.

A specified research question your model addresses:

My model will address whether the survey people's answers to the questions in the data set correlate well with age groups. To do this, I will be testing 4 different types of models and comparing three other models and decide which model projects well to this question.

Survey Questions I will be using for the dataset?

Age vs. These Questions

Have you ever tried to meet up with hometown friends on Thanksgiving night?

Have you ever attended a Friendsgiving?

Will you shop any Black Friday sales on Thanksgiving Day?

Do you work in retail?

How would you describe where you live?

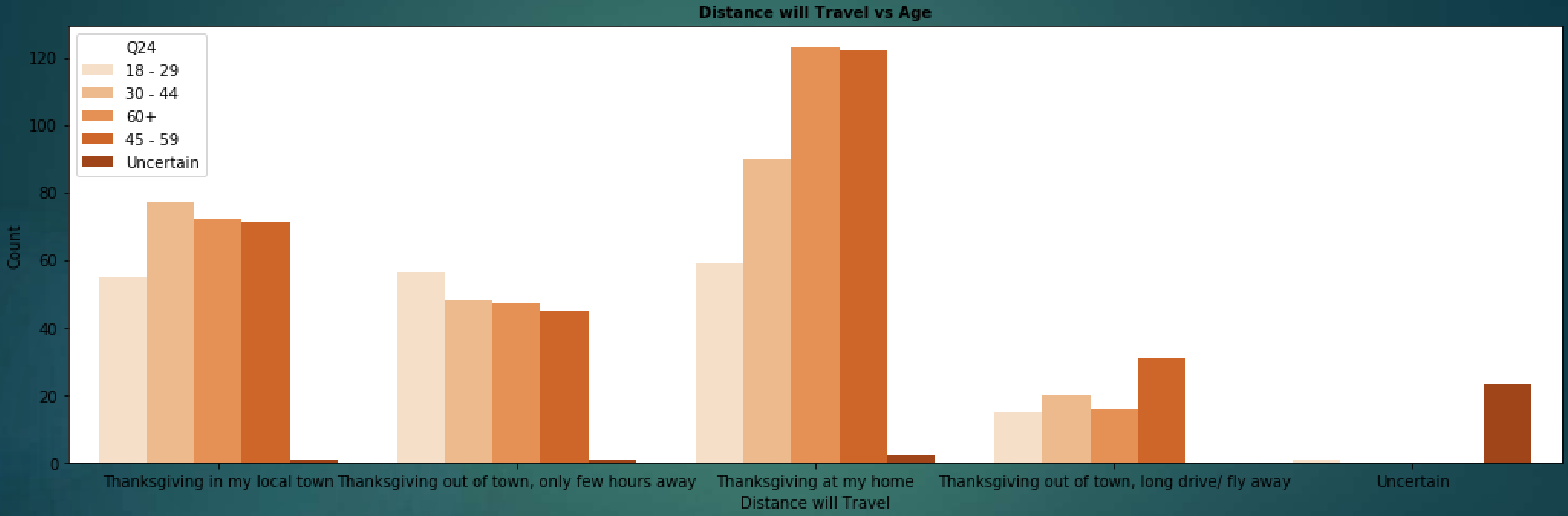
What is your gender?

How much total combined money did all members of your HOUSEHOLD earn last year?

US Region

DATA ANALYSIS

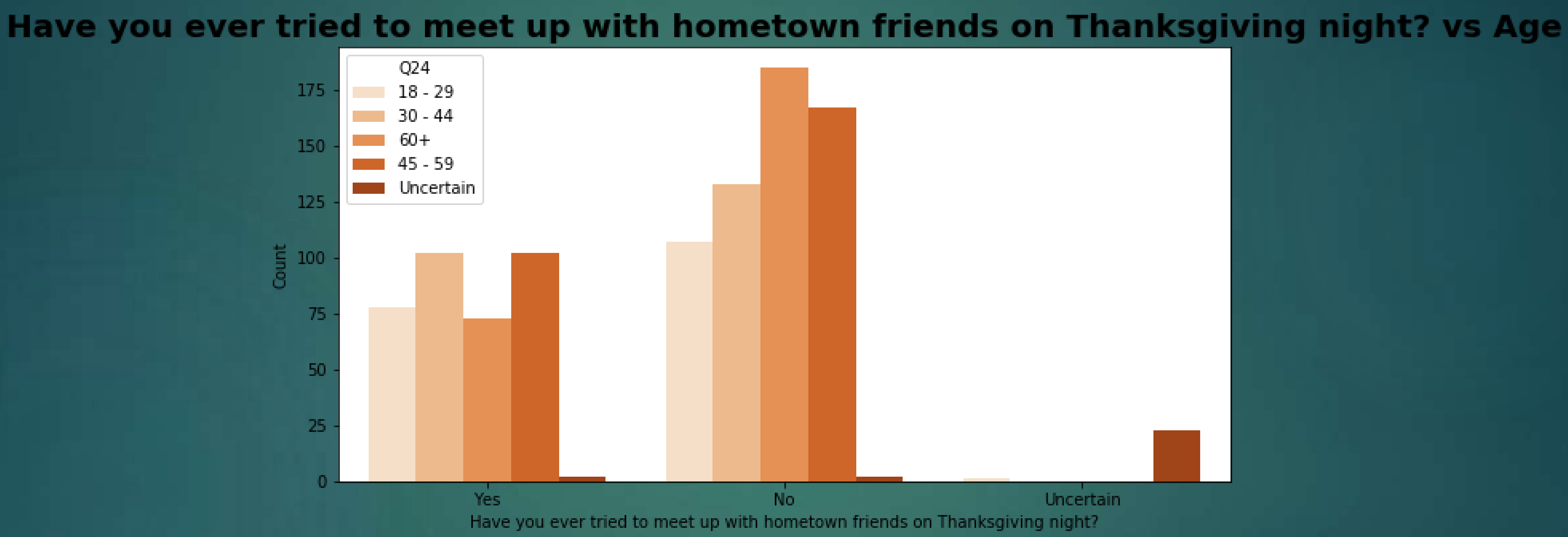
Distance will travel for Thanksgiving?



Distance will Travel vs Age

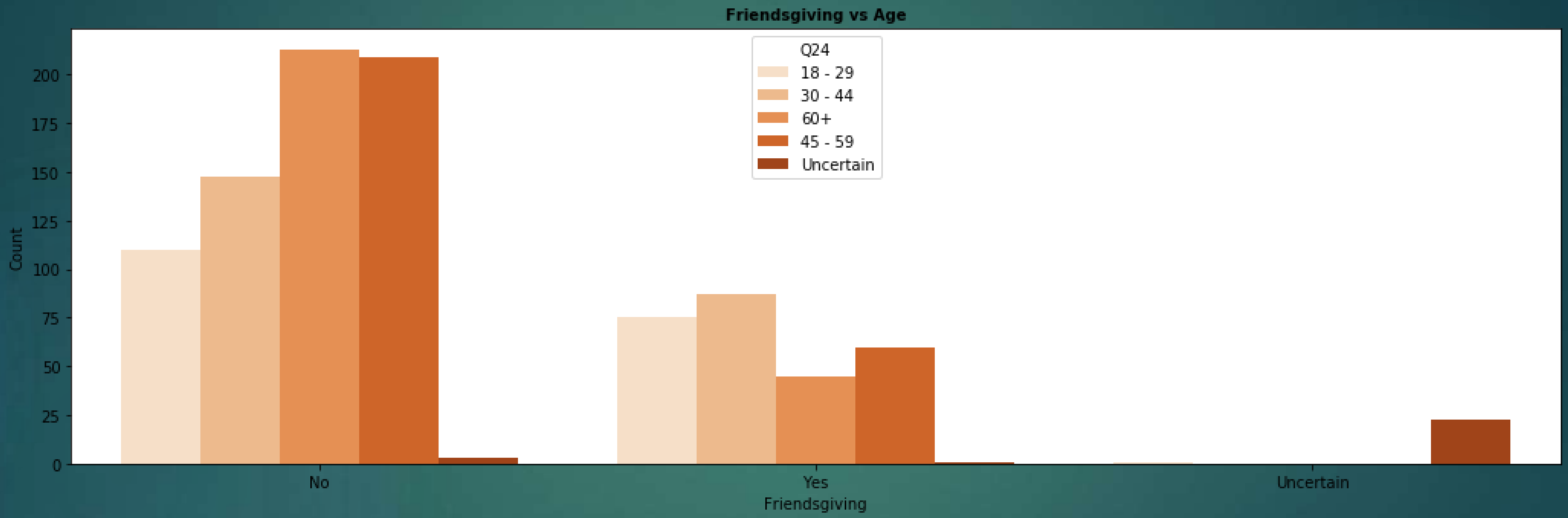
- More millennials like to stay more at home for Thanksgiving and will not go out of their way to drive many hours for the holiday weekend.

Meet with Friends on Thanksgiving?



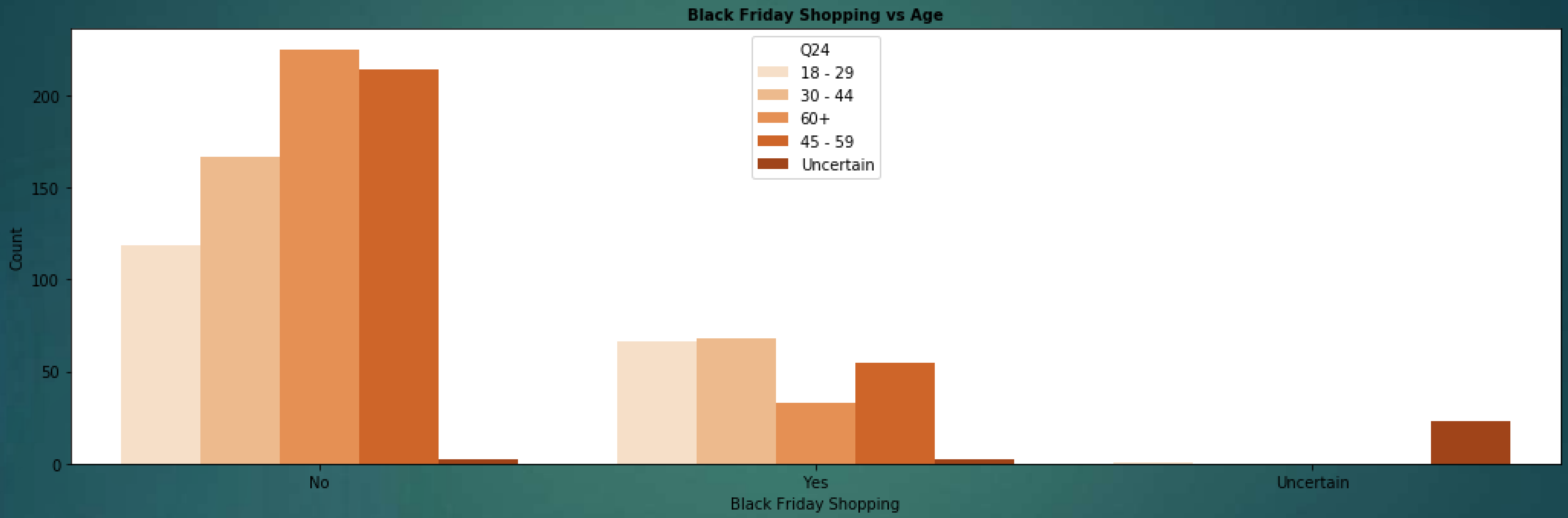
- More of the millennials like to meet up with friends during the holidays (to those who responded to Yes).

Attended a Friendsgiving?



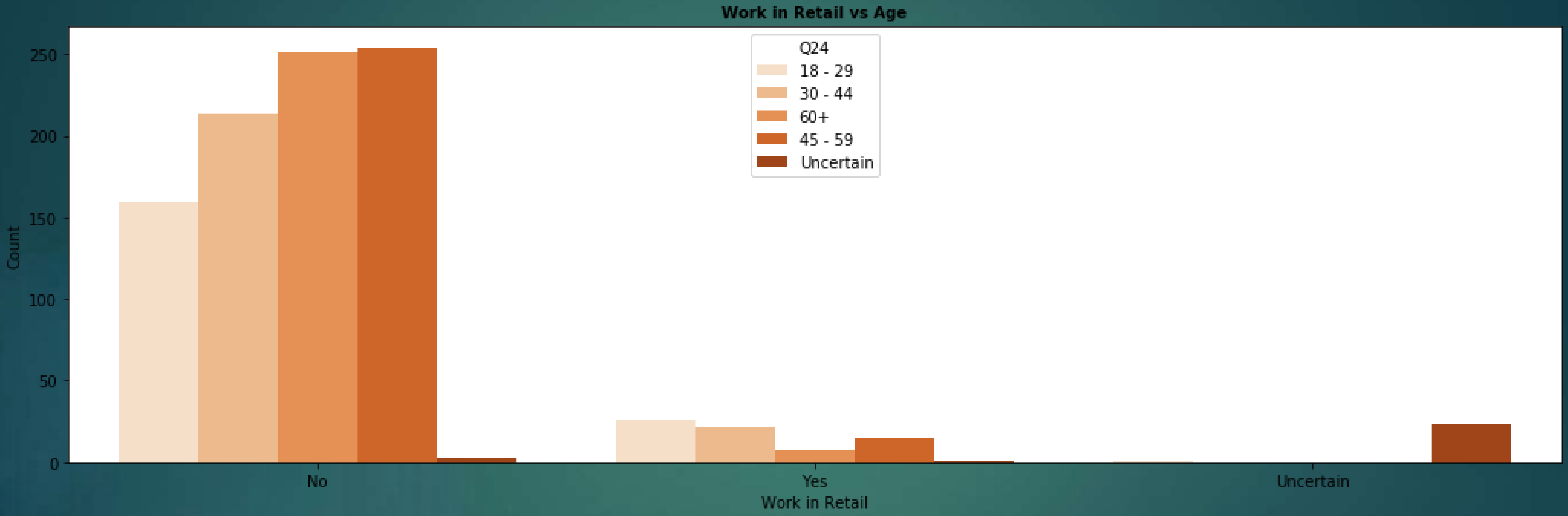
- To those who voted yes, the age of 45 or younger attended a Friendsgiving with their friends.

Do you Black Friday Shopping?



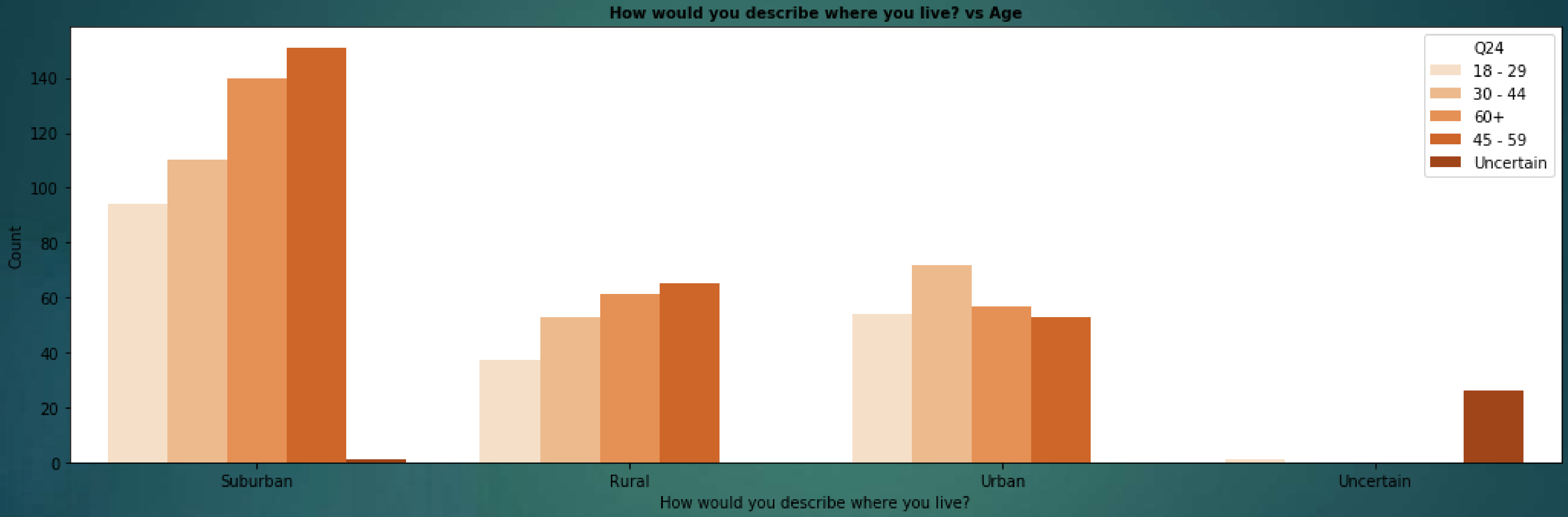
- To those who voted yes, the age of 45 or younger enjoy Black Friday shopping.

Do you work in retail?



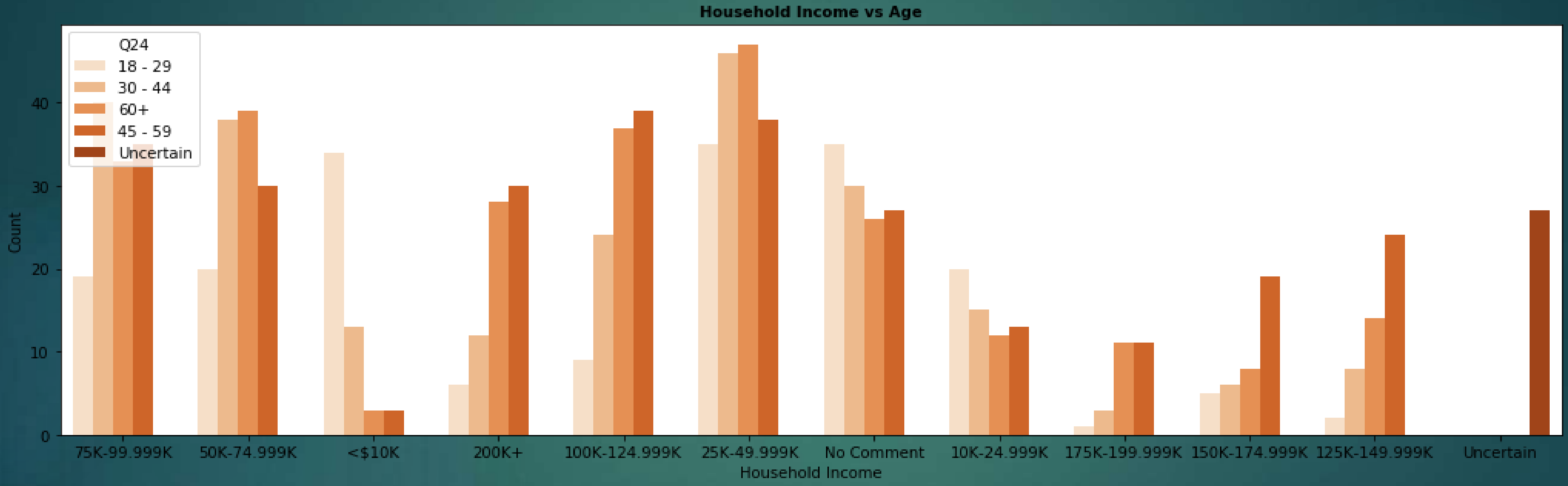
- Most people in this survey do not work in retail especially older adults. However, to those who voted yes, it's no surprise most millennials work in the retail business.

Describe Where You Live



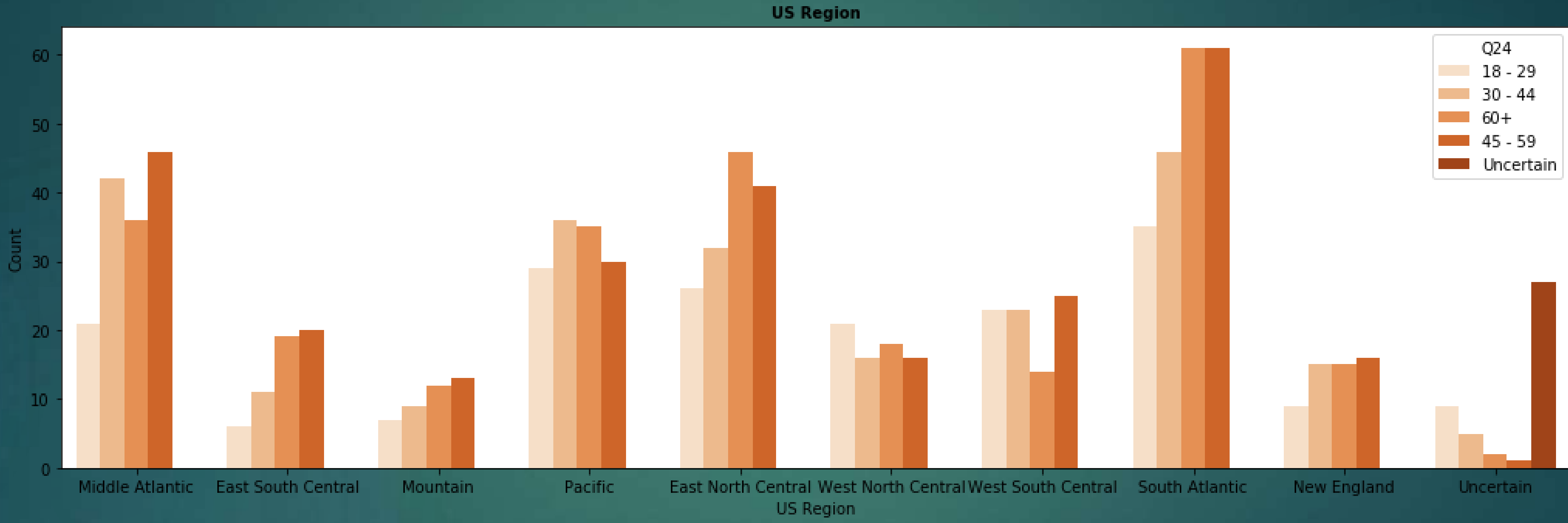
- Based on the survey, most people live in the suburban area based on this surveys. With Suburban and Rural, the age groups got older in both groups.

Household Income



- The majority of the people who took this survey make a modest amount of money on a yearly basis.

US Region



- Based on this category, the regions are diverse. More older adults represent better in this category.

Prediction Models

Analysis

How you chose your model specification and what alternatives you compared it to:

I chose Logistic Regression because it was straightforward to use and easy to train. I chose Gradient Boosting because of the decision trees that can predict which age group has the most popularity with each of the questions. Also, it handles null values, which may be the most useful model for my dataset. I chose Random Forest because it is faster to produce results. Also, my data is not all balanced so this feature is efficient to use. I chose Support Vector Classifier due to its flexibility for datasets.

Gradient Boosting, Random Forest, Support Vector Classifier, and Logistic Regression were the models I chose for this based on the lessons I read. I checked on all those 4 models to determine which model would fit accurately with my model.

Support Vector Classifier had the lowest accuracy score of 32.76% and the lowest overall for cross-validation scores. Gradient Boosting had the second lowest accuracy score of 34.81%. Random Forest had the second highest accuracy score of 35.15%. Logistic Regression had the highest accuracy score with 37.15%.

The cross-validation scores were not consistent with the accuracy scores because Gradient Boosting had higher cross-validation scores than Random Forest and Logistic Regression despite it was the second lowest accuracy score.

Results/Conclusion

The practical uses of your model for an audience of interest:

The purpose of the dataset was to see how everything correlated together and it seems most young people who try to meet up with other friends on Thanksgiving enjoy Black Friday shopping and those type of people work in the retail business. These surveys can inspire other people to try out other activities during that holiday season.

Any weak points or shortcomings of your model:

The correlations in the variables (even after trying to drop to variables) did not correlate high with all of the other ones. The accuracy rates may have gone higher if some of the respondents actually were honest during the survey (only 1 respondent said they did not celebrate Thanksgiving, for example). However, because there was four different age groups, there was a 25% chance that someone would have a chance of guessing the age group accurately. Therefore, my accuracy rates for those four models were not as bad as I anticipated.