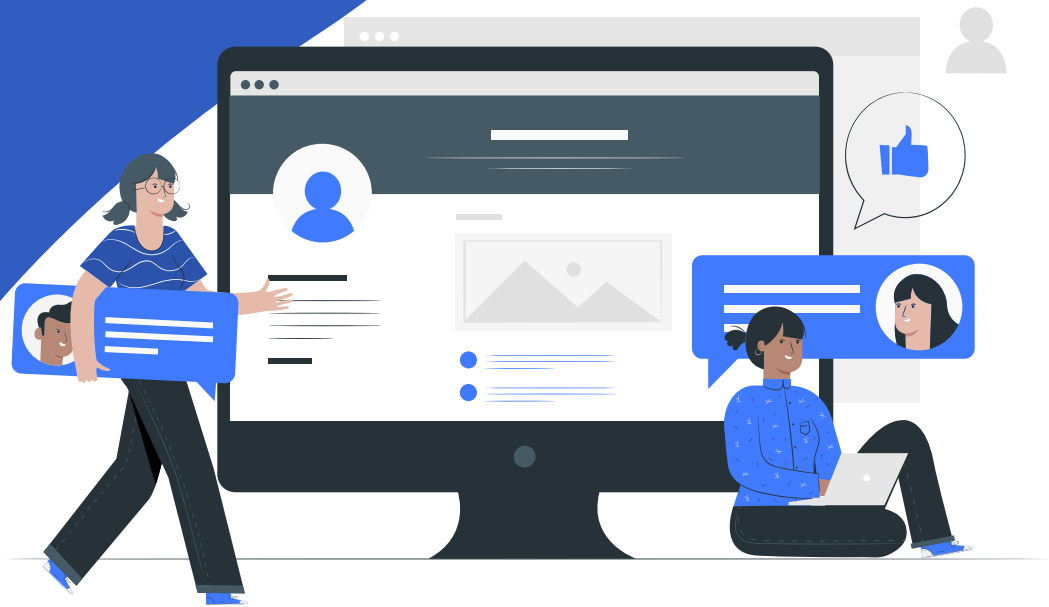
Deloitte Survey Analysis



- Team HyperNet
  Ruchi, Yuanmo (Caroline), Maxine

Our team

Ruchi Bhatia
Data Analyst, HyperNet

Caroline Zhu
Data Analyst, HyperNet

Maxine Ma
Data Analyst, HyperNet

Agenda

Problem Statement

What problem is being solved?
Who benefits from the problem solution?

EDA + Data Preparation

What does the data look like?
How was the data prepared for modeling?

Models + Results Evaluation

Supervised & Unsupervised Learning
Model choice & Results evaluation

Recommendation

Recommendation & next steps

What are we aiming to solve?

Who:

HyperNet - Home internet service provider

What:

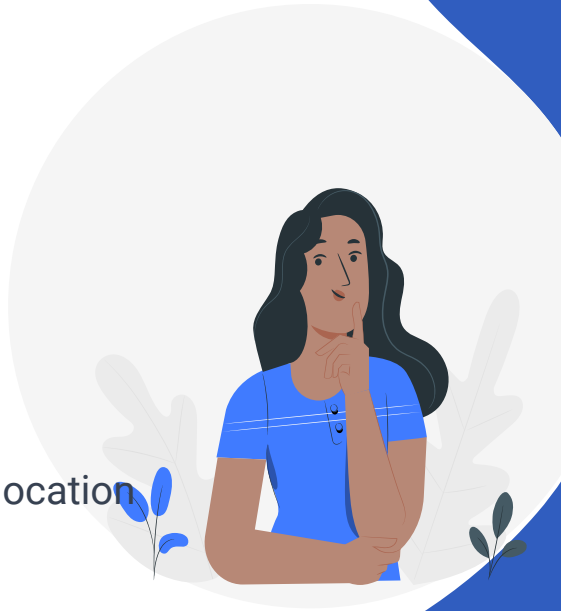
Identify the demand for the internet speed upgrade option in the market

What are we aiming to solve?

Why:

To make informed decisions that can lead to **revenue growth** and

increased customer satisfaction and loyalty



How:

- Understanding the reasons behind customers' interest, and
- Identifying patterns in customer demand based on demographics/location

What Does Our Data Look Like ?

2131 Rows X 196 Columns

Each row represents 1 survey answer

Each column represents:
- either all the choices of 1 question or
- a binary choice of 1 option of the question

Questions include:

- Demographic Questions: Age/Gender/Employment Status, etc.
- Media Owned or Planned to Owned Questions
- Media Value Ranked Questions
- Time Spent Preference Questions
- Media subscription
- Entertainment habits

EDA & Data preparation

Step 1

Filter the data based on the problem statement

Step 2

Demographic analytics

Age, Gender, Region, Income, etc

Step 3

Handle missing values

Drop columns with many nulls

Fill rest nulls based on questions

Step 4

Summary Statistics

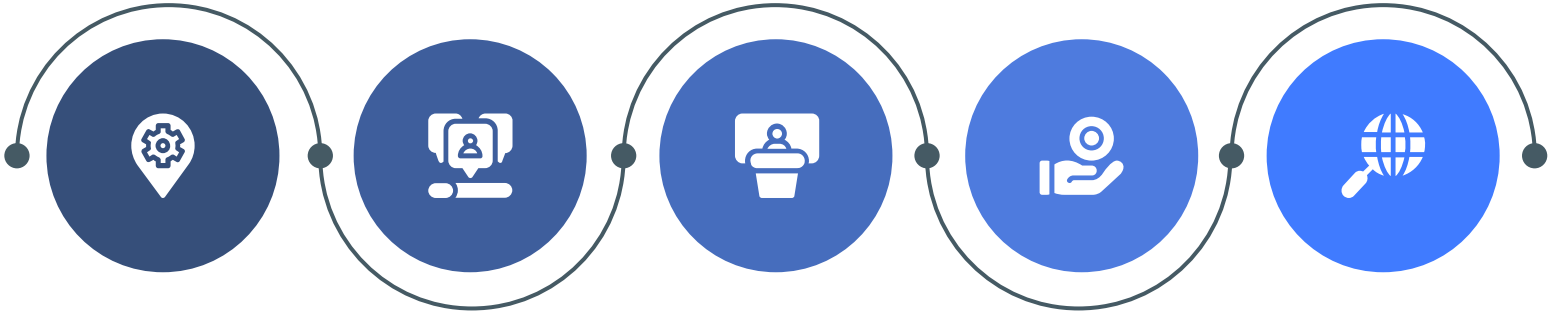
Univariate Analysis

Bivariate Analysis

Step 5

Encoding categorical data

Binary Encoding, Label Encoding, One Hot Encoding



Step 1: Filter the data based on the problem statement

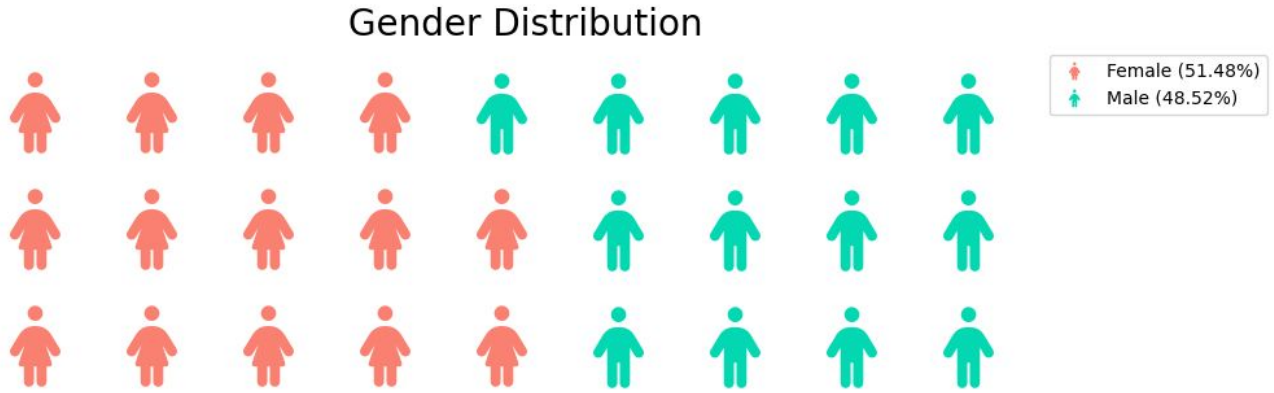
- To figure out who are willing to pay more for higher Internet speed, we chose the Q29 as our target variable:

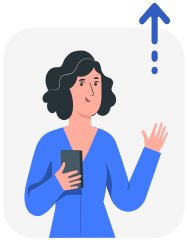
Q29 - You said that you subscribe to home Internet access, how much more would you be willing to pay to receive double your download speed?

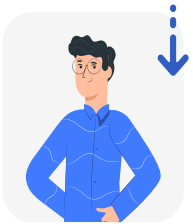
```
# Filter the data frame
internet_df = data[data['Q26 - Which of the following subscriptions does your household purchase?-Home internet']
                   == 'Yes']
```

- To solve our problem, we first use **Q26: whether or not owning home Internet access** as a filter

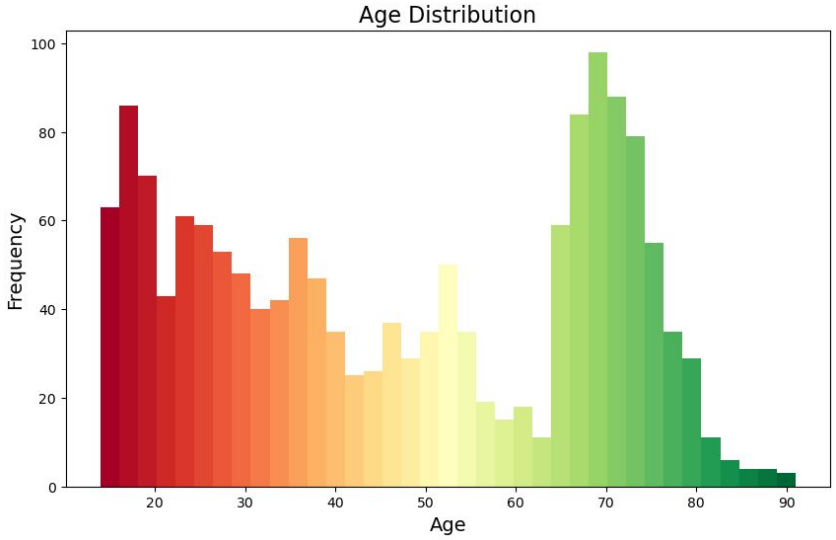
Step 2: Demographic of Target Customers

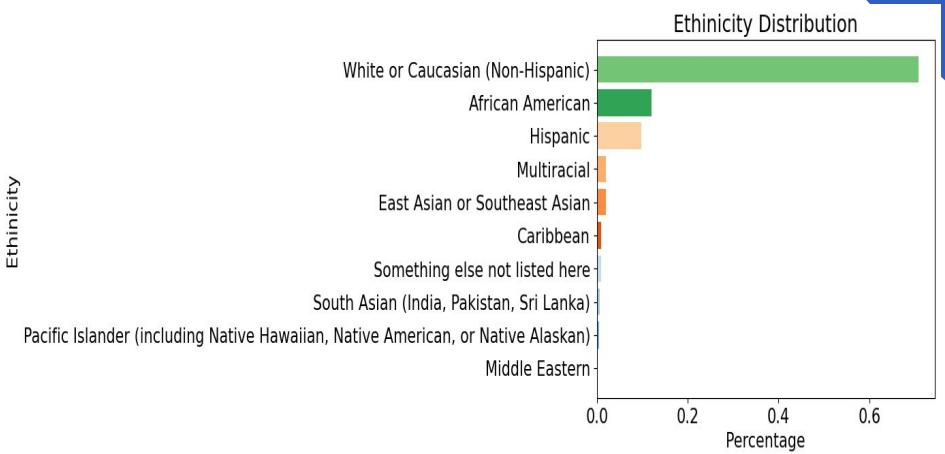




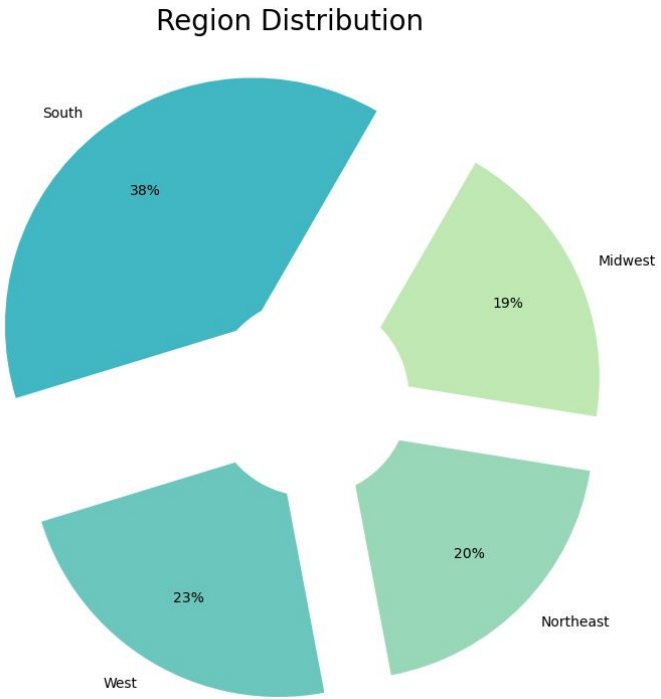


Step 2: Demographic of Target Customers

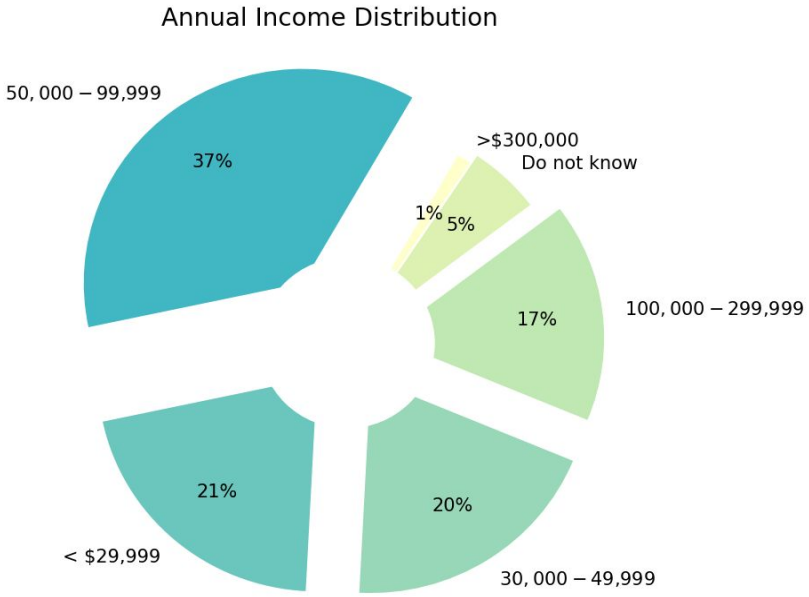


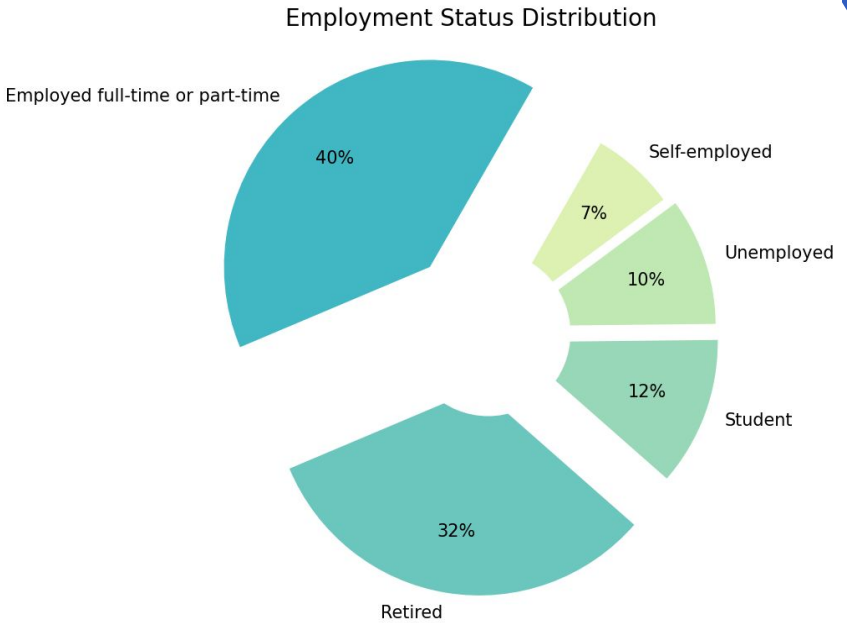


Step 2: Demographic of Target Customers



Step 2: Demographic of Target Customers





Step 3: Handle Missing Values

Drop columns with too many nulls

with missing values > 60%

Drop columns with limited information

Record number
Final weights

Fill rest nulls based on questions

Replace NA with 4 in rank questions
Replace NA with -999 in cells as not answered

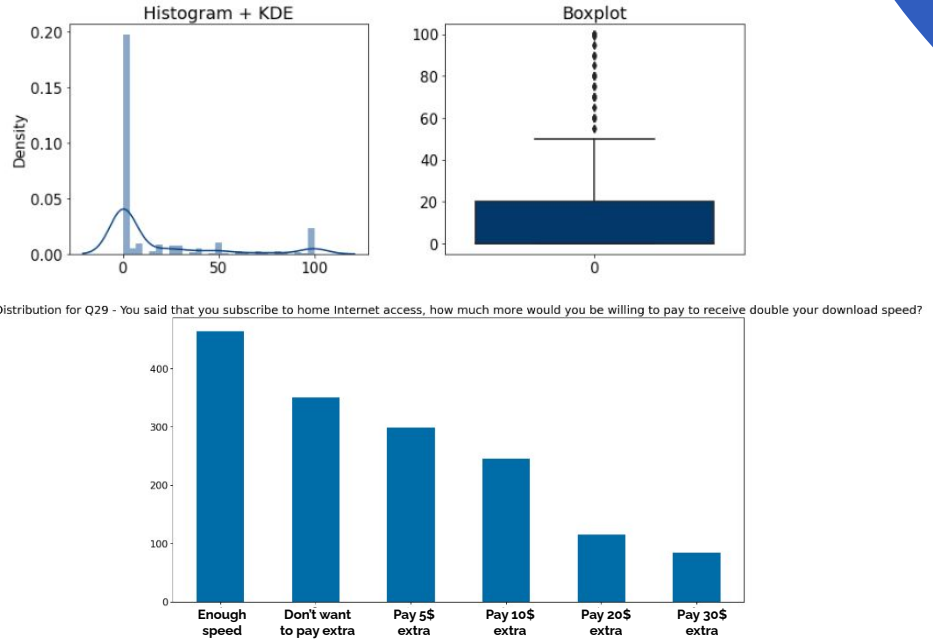
Step 4: Summary Statistics & variate analysis

Divide dataset into **Categorical-only & Numerical-only** datasets
To perform summary statistics, univariate analysis

	Q4 - What is your gender?	age - you are...	Q2 - In which state do you currently reside?	region - Region
count	1558	1558	1558	1558
unique	2	6	50	4
top	Female	70 or older	California	South
freq	802	369	168	593

	Q1r1 - To begin, what is your age?	Q15r1 - Smartphone - Of the time you spend watching movies, what percentage of time do you watch on the following devices?	Q15r2 - Tablet - Of the time you spend watching movies, what percentage of time do you watch on the following devices?
count	1558.000000	1558.000000	1558.000000
mean	47.308087	9.435815	7.626444
std	21.493291	19.208739	17.758325
min	14.000000	0.000000	0.000000
25%	27.000000	0.000000	0.000000
50%	47.000000	0.000000	0.000000
75%	69.000000	10.000000	5.000000
max	91.000000	100.000000	100.000000

Summary Statistics



Univariate Analysis

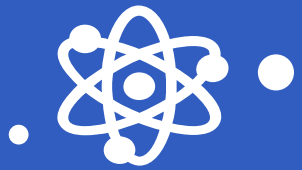
Step 5: Encoding Categorical Data

Binary Encoding

Replace “Yes” or “No” with 1 and 0

Label Encoding

Labelize our target variable Q29 with 1 and 0 based on answer choice:
r1 - r4 : 1 → Willing to pay
r5 - r6: 0 → Unwilling to pay



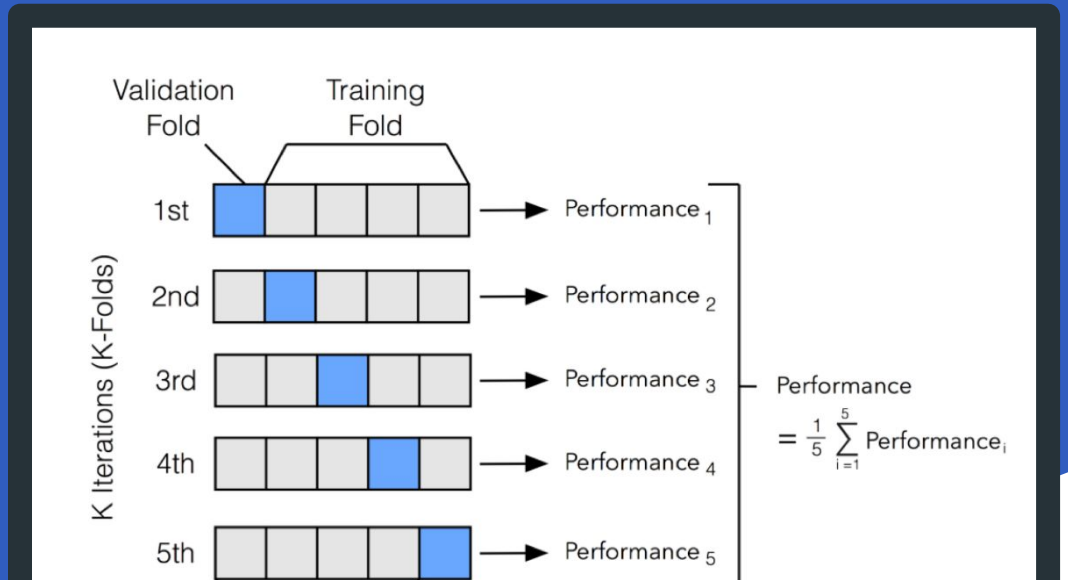
One Hot Encoding

A binary representation of each category
Each category → a column
Each row → a 1 or 0 depending on the category it belongs to

Modeling Approach

Supervised Learning

- Implemented **K-fold cross-validation** to evaluate the performance of **10 different machine learning models**



Model Evaluation

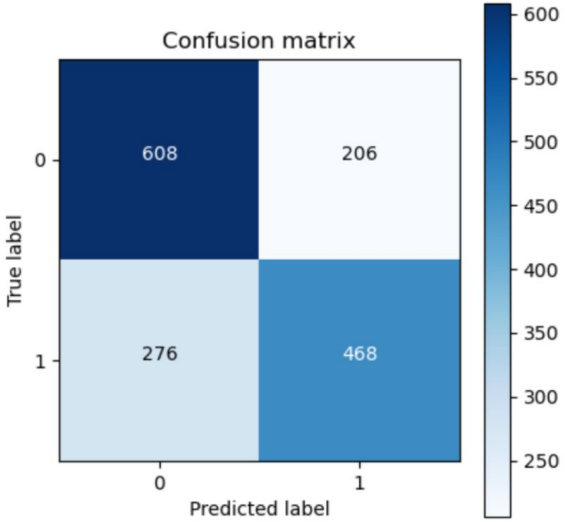
Accuracy

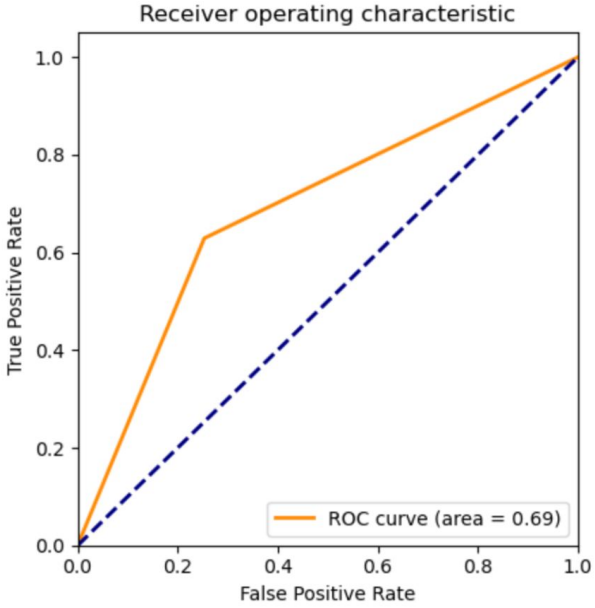
	Training Set	Validation Set
Naive Bayes	0.7439	0.7195
Logistic Regression	0.7042	0.6906
KNN	0.7865	0.7002
SVM	0.5738	0.5738
Decision Tree	1.0	0.6662
Bagging Decision Tree	0.9855	0.7163
Boosted Decision Tree	1.0	0.6752
Random Forest	1.0	0.7432
Voting Classification	0.9136	0.7233
Neural Network	0.7867	0.6630

Other models seem to be overfitting!

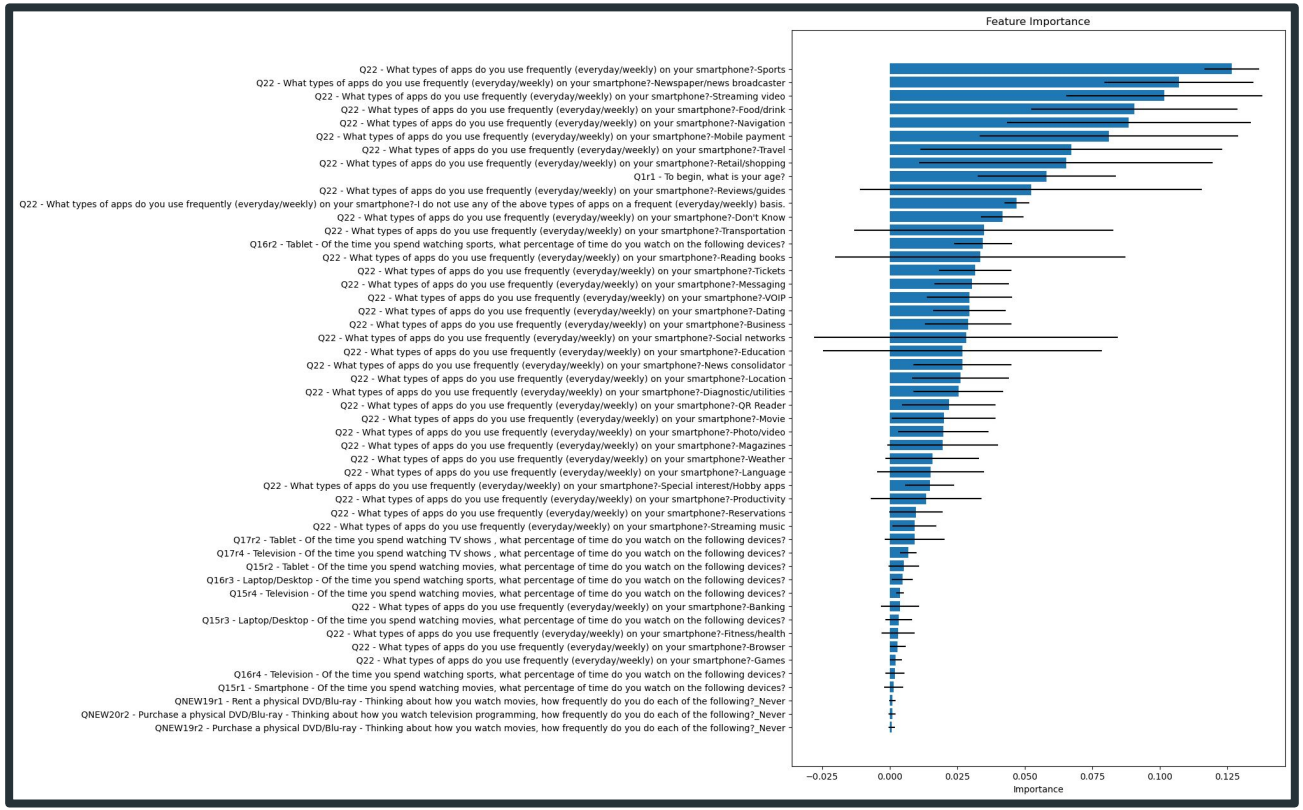
Model Evaluation
Confusion Matrix and ROC

Model: Logistic Regression





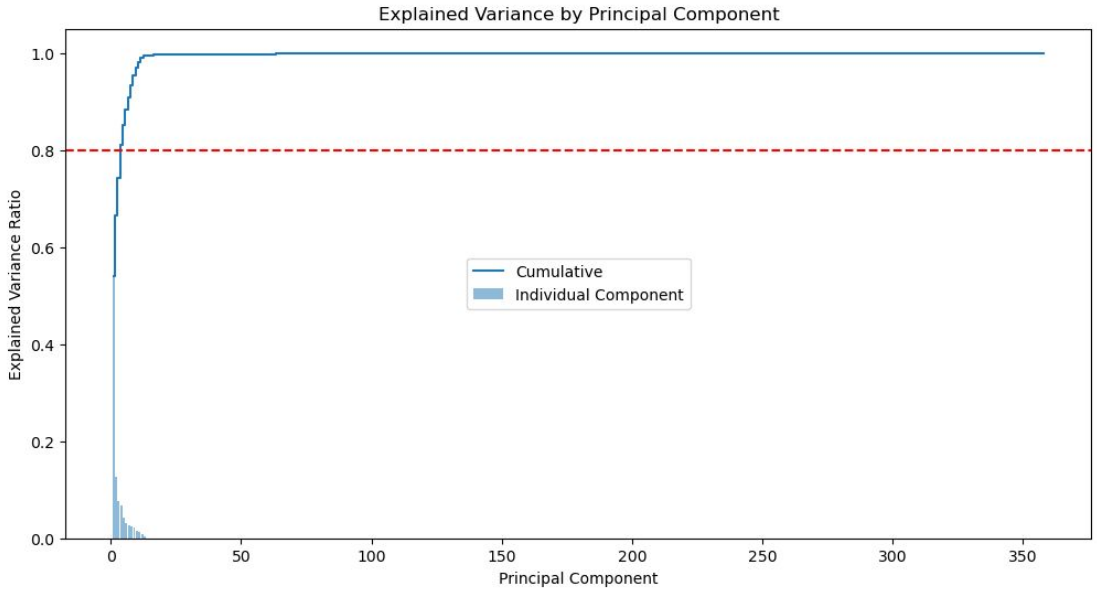
Feature Importance



Modeling Approach

Unsupervised Learning: KMeans

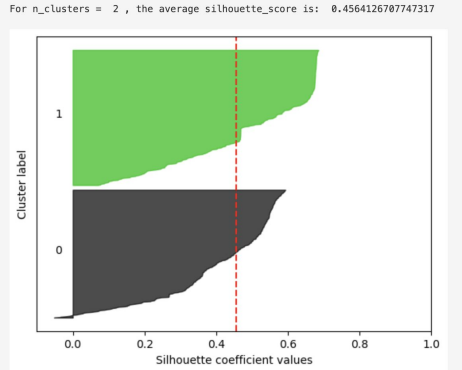
- **PCA**: to reduce the dimensionality of the data

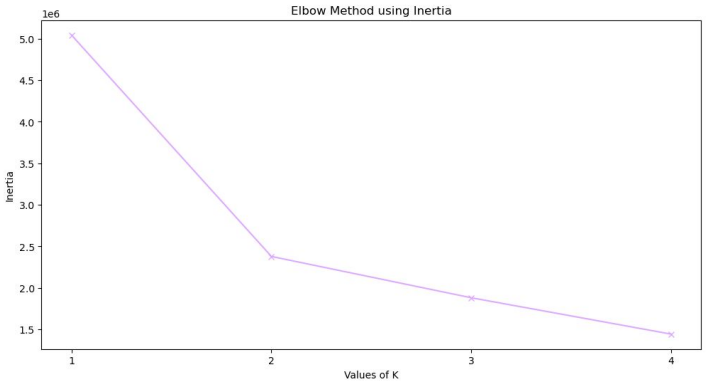


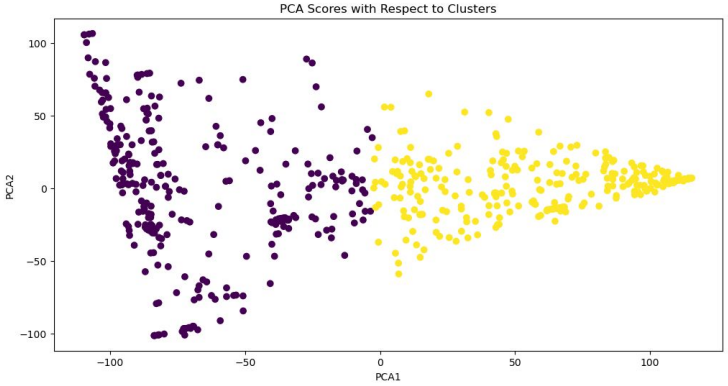
Modeling Approach

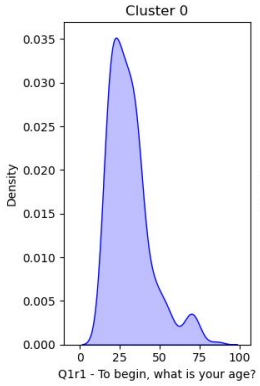
Unsupervised Learning: KMeans

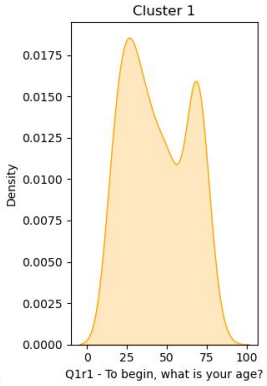
- **KMeans** clustering to group similar data points together based on their proximity in the reduced-dimensional space.







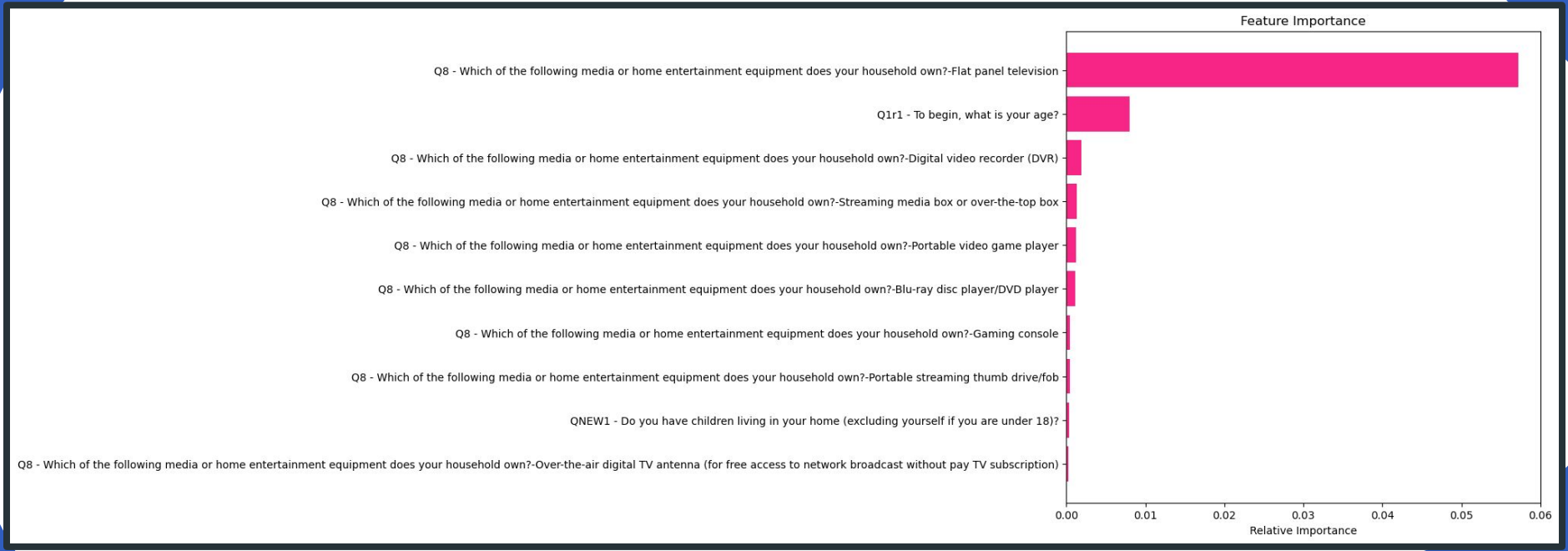




Cluster Distinguishing Features

Unsupervised Learning: KMeans

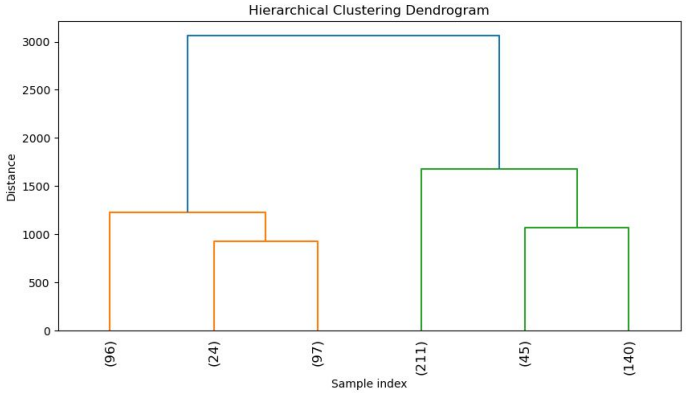
- Trained a feature-selecting classifier (Random Forest) on cluster labels
- Inspected classifier for most important features



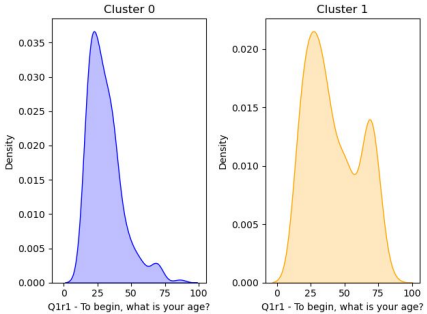
Modeling Approach

Unsupervised Learning: Agglomerative Clustering

- **Dendrogram:** to determine the optimal number of clusters



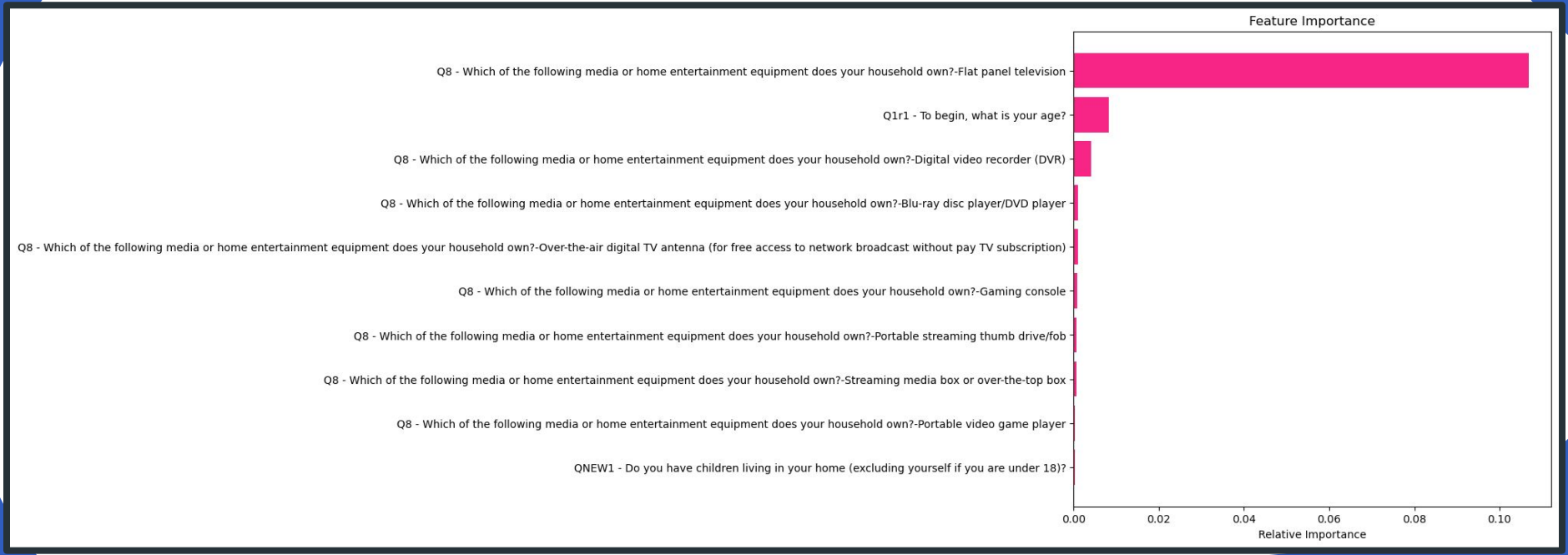
- **Agglomerative Clustering:** to group similar data points together based on their similarity in the original feature space



Cluster Distinguishing Features

Unsupervised Learning: Agglomerative Clustering

- Trained a feature-selecting classifier (Random Forest) on cluster labels
- Inspected classifier for most important features



Recommendations

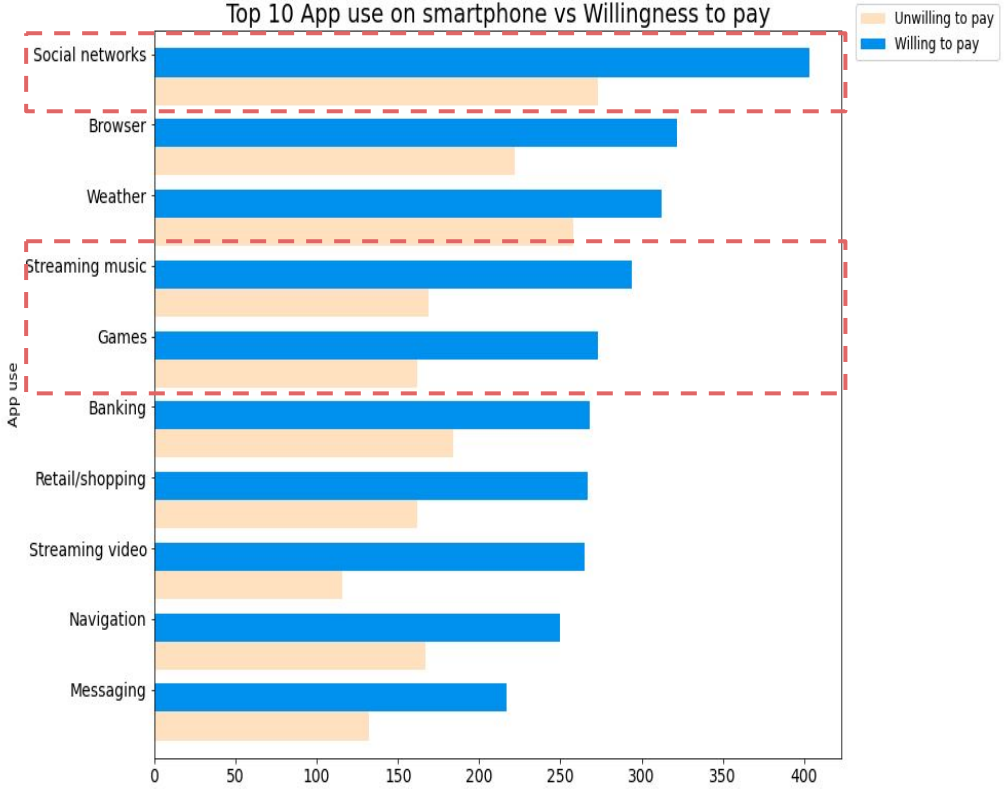
Feature Analysis based on

the **Top 8 variables** with **highest feature importance** scores

in **Logistic Regression**.

- App Use On Smartphones
- Age & Employment
- Region & States
- Income
- Time Spent watching sports on Laptop
- Time spent watching shows & movies on Tablet

App Use On Smartphones



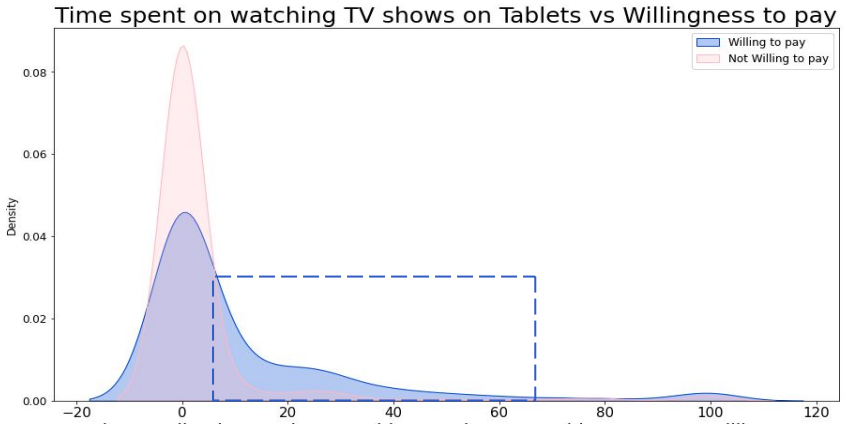
Q22 - What types of apps do you use frequently (everyday/weekly) on your smartphone?

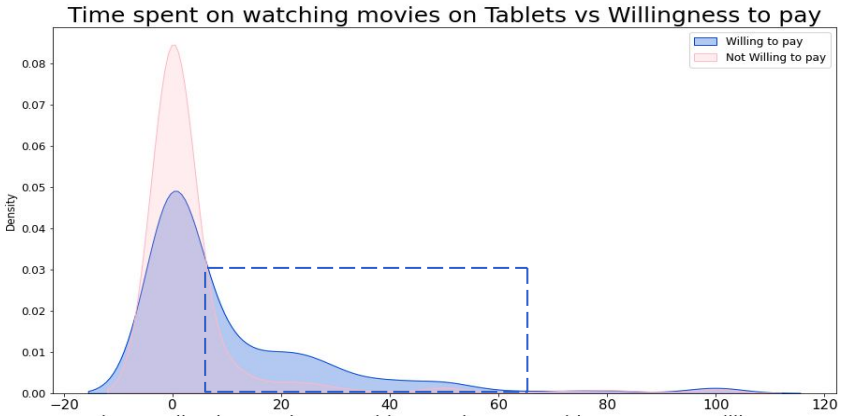
Collaborations can be made with

- **Social media companies**
  Instagram, Facebook, TikTok, Snapchat
- **Streaming music companies**
  Spotify, Apple Music
- **Smartphone Game companies**
  Supercell, Niantic, Innersloth

⇒ **Offer customized Internet bundle services(e.g. 10 GB for listening to music on Spotify).**

Time spent watching shows & movies on Tablet





Q15r2 - Tablet - Of the time you spend watching movies, what percentage of time do you watch on the following devices?
-TV shows -movies

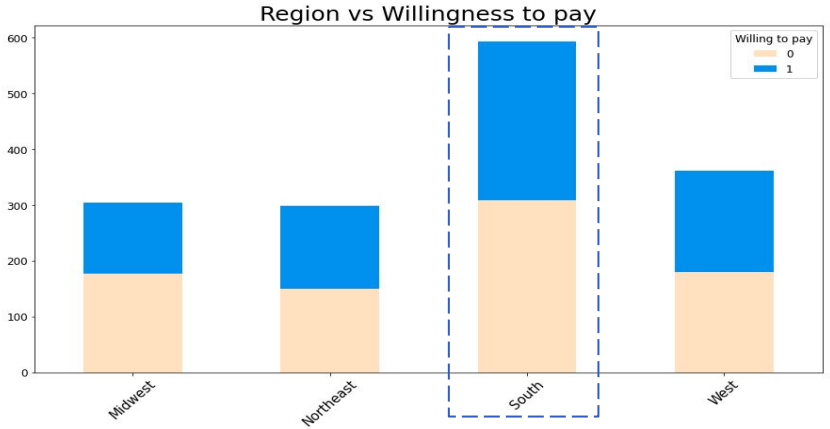
Collaborations can be made with

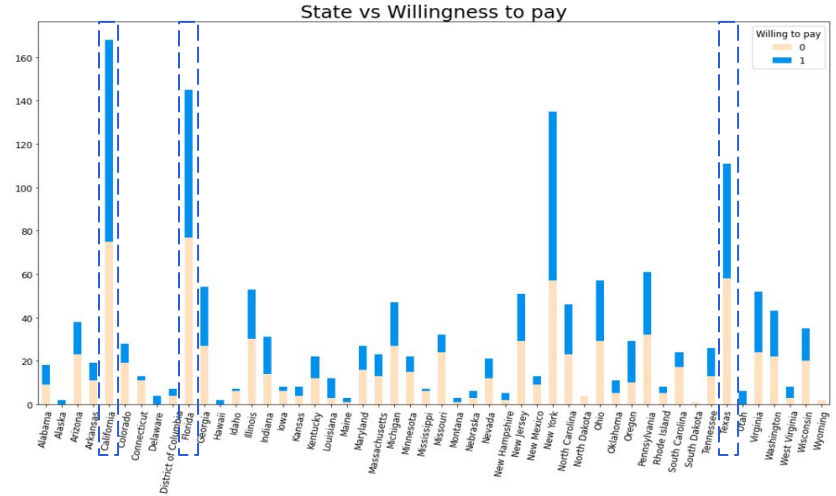
- **Streaming video companies**
  Youtube, Netflix, fuboTV, Disney+, Amazon Prime Video

⇒ **Offer customized Internet bundle services.**

⇒ **Sell Internet packages with video platform memberships.**

Region & States





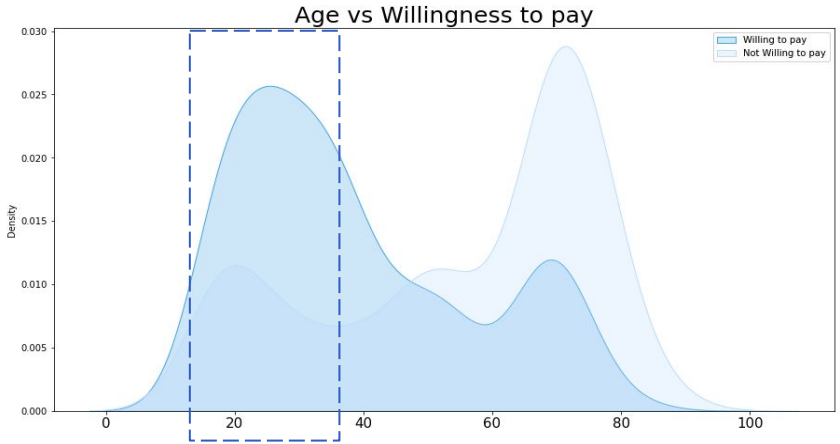
region - Region

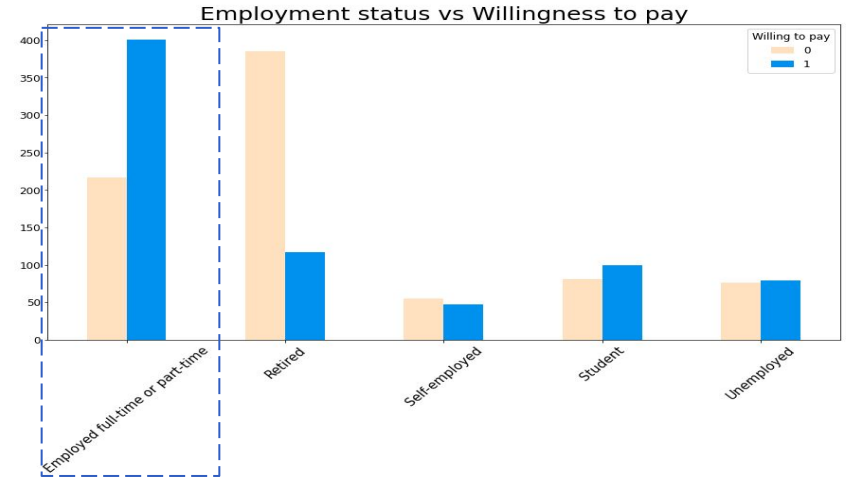
Q2 - In which state do you currently reside?

Further expansion of business will mainly focus on

- **Southern states**
  California, Florida, Texas.

Age & Employment





Q1r1 - To begin, what is your age?

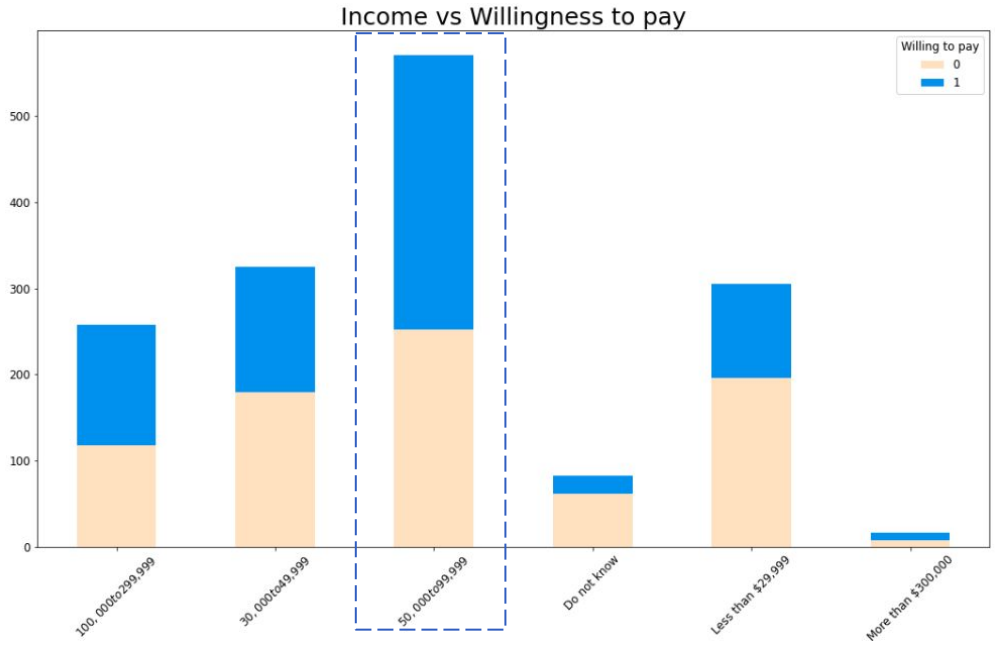
QNEW3 - What is your employment status?

When making sales calls or sending sales emails, mainly Target customers who are

- **Young adults**
  with a rough age range [16,38]

- **Employed full-time or part-time**
  since these customers are gaining income to pay.

Income

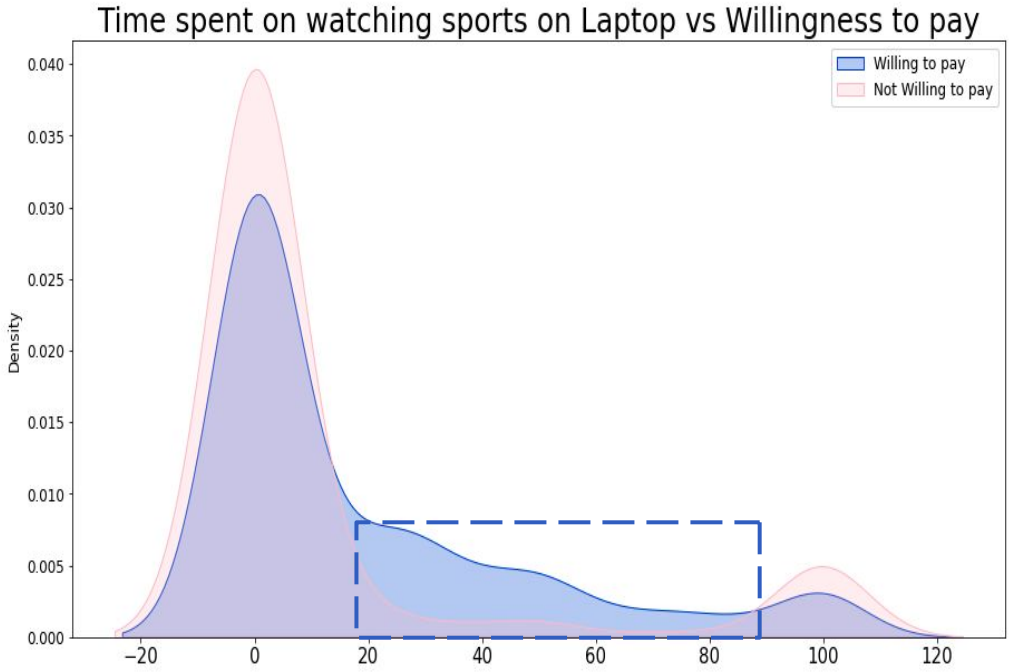


Q6 - Into which of the following categories does your total annual household income fall before taxes?

Target mainly households that are

- **Middle class households**
  With an annual households income range [50k,99k] $

Time Spent watching sports on Laptop



Q17r2 - Tablet - Of the time you spend watching TV shows , what percentage of time do you watch on the following devices?

Offer seasonal Internet surfing service for

- **Sporting events**
  Super Bowl, NBA Finals, Iron Man

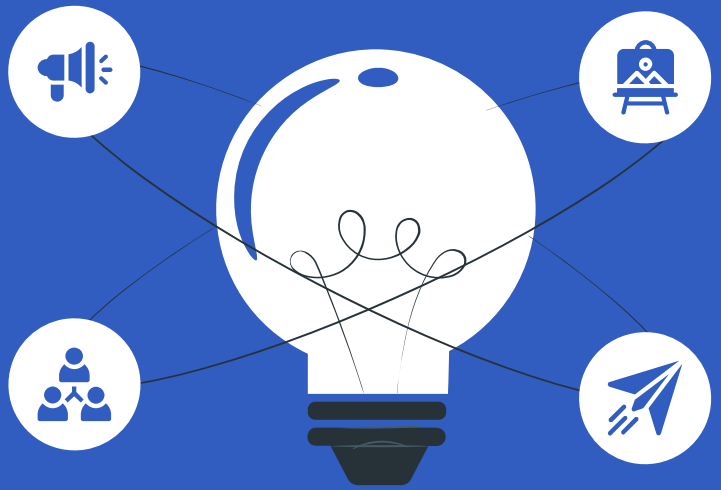
Summary of Recommendations

Collaborations

With online entertainment companies to offer customized Internet bundle services

Target Personas

Young Adults employed full-time or part-time, Middle class households



Business Expansion

Focus on southern states as higher willingness to pay

Seasonal offers

Offer seasonal Internet surfing service for highly popular sports events

Thanks!

Hypernet: Get Connected to the World in a Flash!