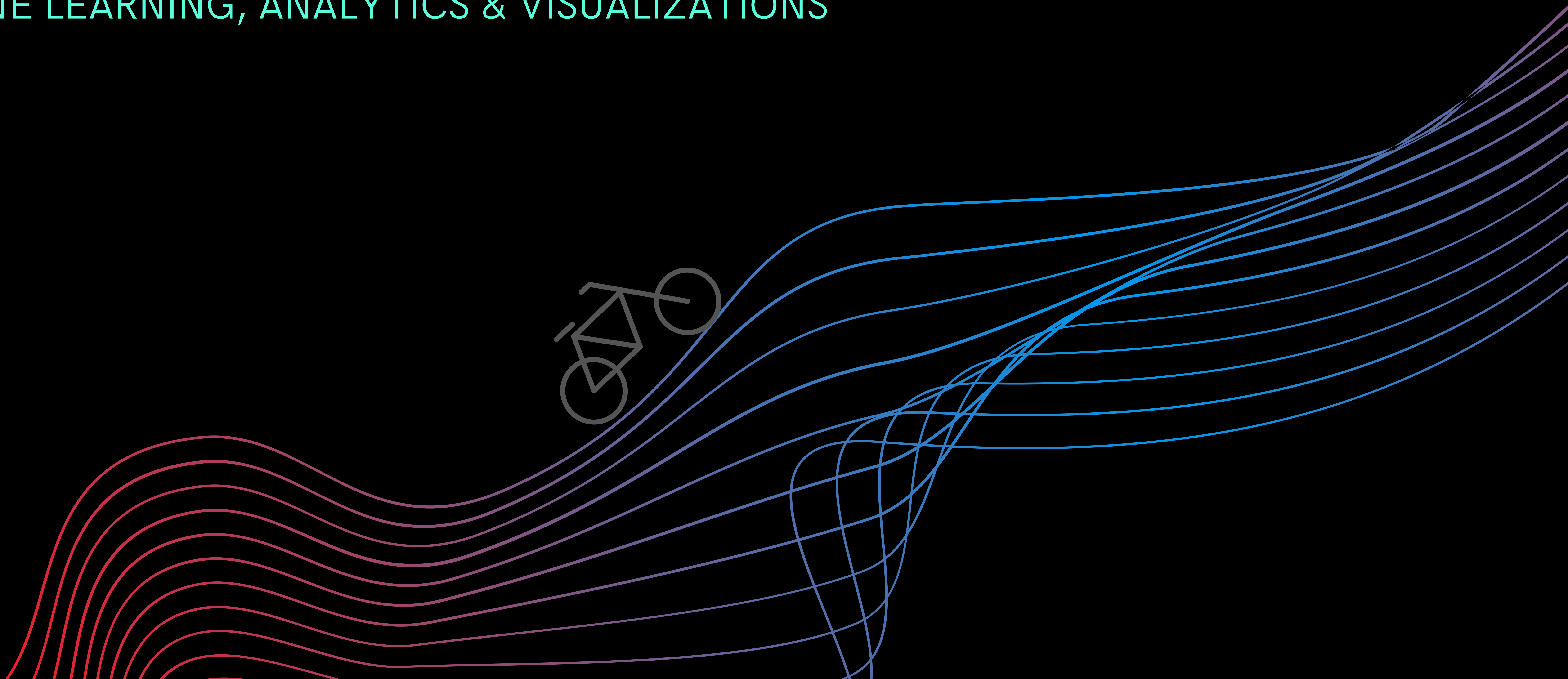
NYC citibike

MACHINE LEARNING, ANALYTICS & VISUALIZATIONS



WE ARE DISRUPTORS AND INNOVATORS. ARE WE?...

We work at Citi, our team handles performance analytics for non-digital marketing campaigns...

...after 6 years of Citibike evolving, we took the initiative to seek a better alignment between Citi goals and riders...

...our market reach and visibility are the perfect mix to shorten the existing gap...

I AM SURE YOU KNOW, BUT JUST IN CASE...

Citi Bike is New York City's bike share system. Citi Bike launched in May 2013 and has become an essential part of our transportation network.

Citi Bike, consists of a fleet of bikes that are locked into a network of docking stations throughout the city. The bikes can be unlocked from one station and returned to any other station in the system.

People use bike share to commute to work or school, run errands, get to appointments or social engagements, and more.

Citi Bike is available for use 24 hours/day, 7 days/week, 365 days/year.

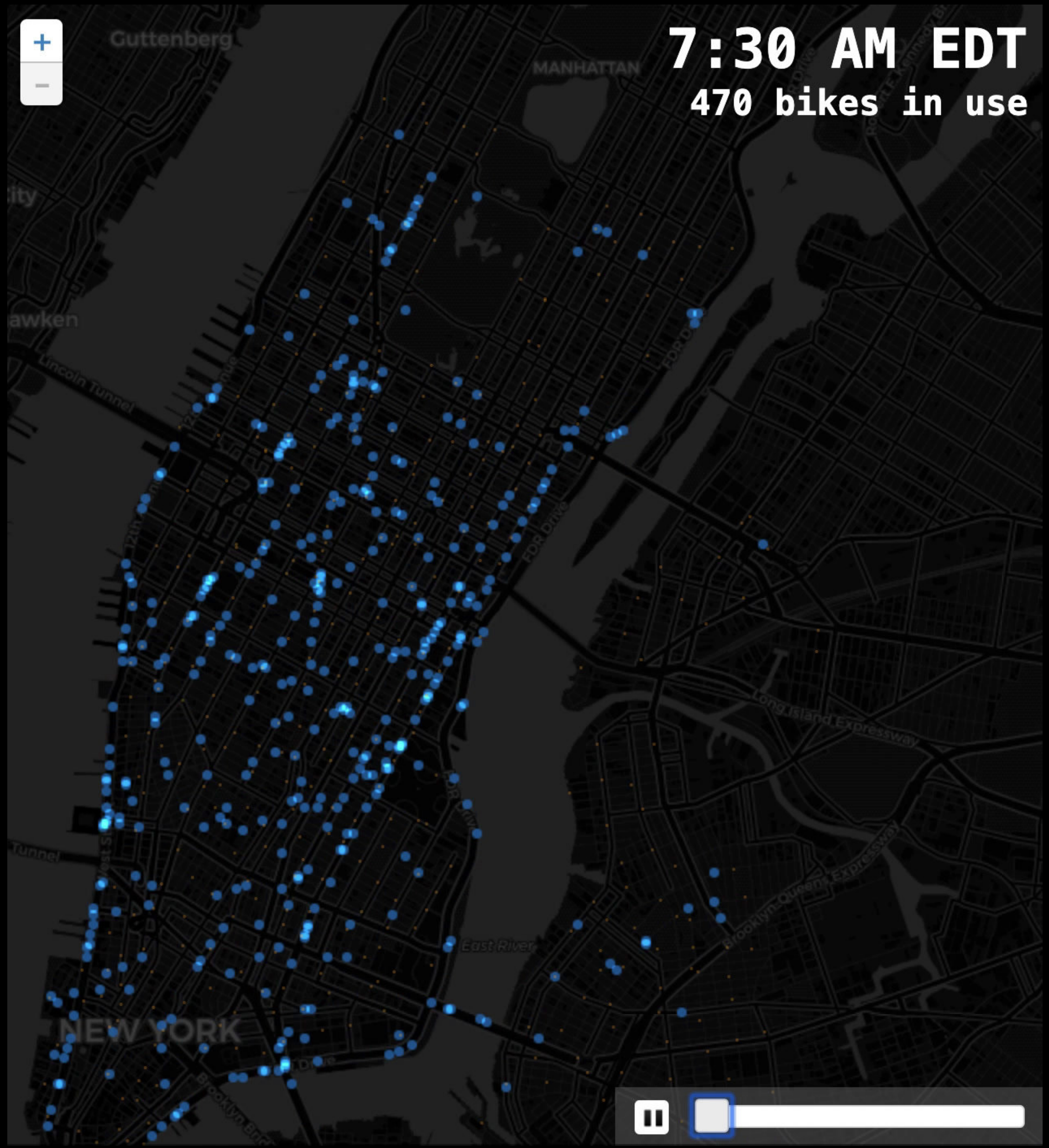
ABOUT THE DATA

- There are currently about 12,000 shared bikes and 750 docking stations.

- The immense available usage data can be used to seek trends and insights.

- For each month since the system's inception, there is a file containing trip details.

- We have decided to use the data from every September from 2013-2019.



ABOUT THE DATA

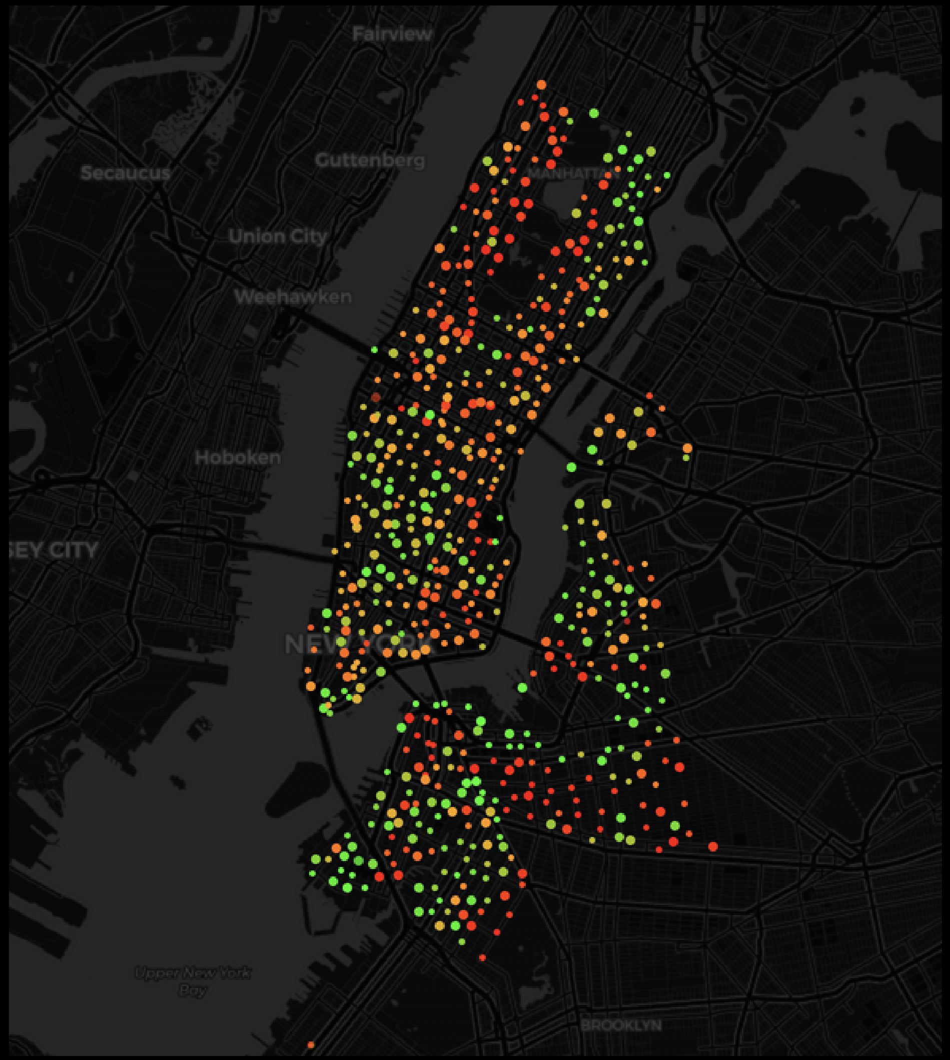
Attribute	Metadata of Attribute
Trip Duration (in seconds)	Shows the time it took to finish the bike ride
Start Trip Timestamp	Timestamp when the trip was started
Stop Trip Timestamp	Timestamp when trip was ended
Start Station ID	Station ID of the station from where the trip was started
Start Station Name	NYC addresses consisting of St and Ave
Start Station Latitude / Longitude	Coordinates of the Start Station Name
End Station ID	station ID of the station where trip ended
End Station Name	NYC addresses consisting of St and Ave
End Station Latitude / Longitude	Coordinates of the End Station Name
Bike ID	Unique ID for the bike
User Type	Customer = 24-hour pass or 3-day pass | Subscriber = Annual Membershi
Gender	(Zero = Unknown; 1 = Male; 2 = Female)
Year of Birth	Birthyear of the user
Age	Age of user

OUR HYPOTHESIS

- By using the location of the bike stations we can predict the user gender and age-group.

OUR OBJECTIVE

- Increase female Citi bike riders population.
- Increase college-age (17-25) riders population.

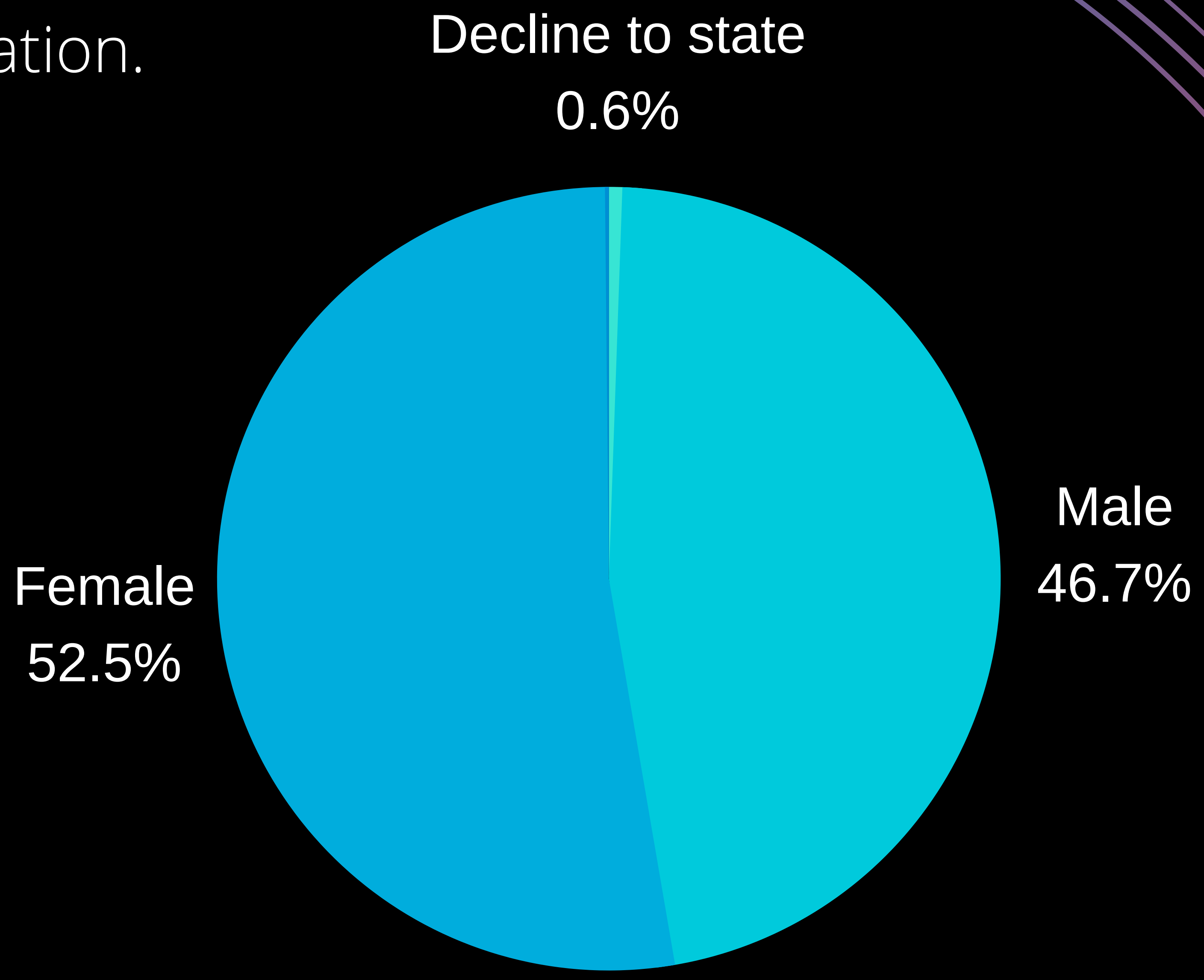


WHY

- NYC population 54 percent female, but male Citi Bike riders are 2.4-4:1 ratio.
- High school and college age students are the second to last age group in terms of ridership.

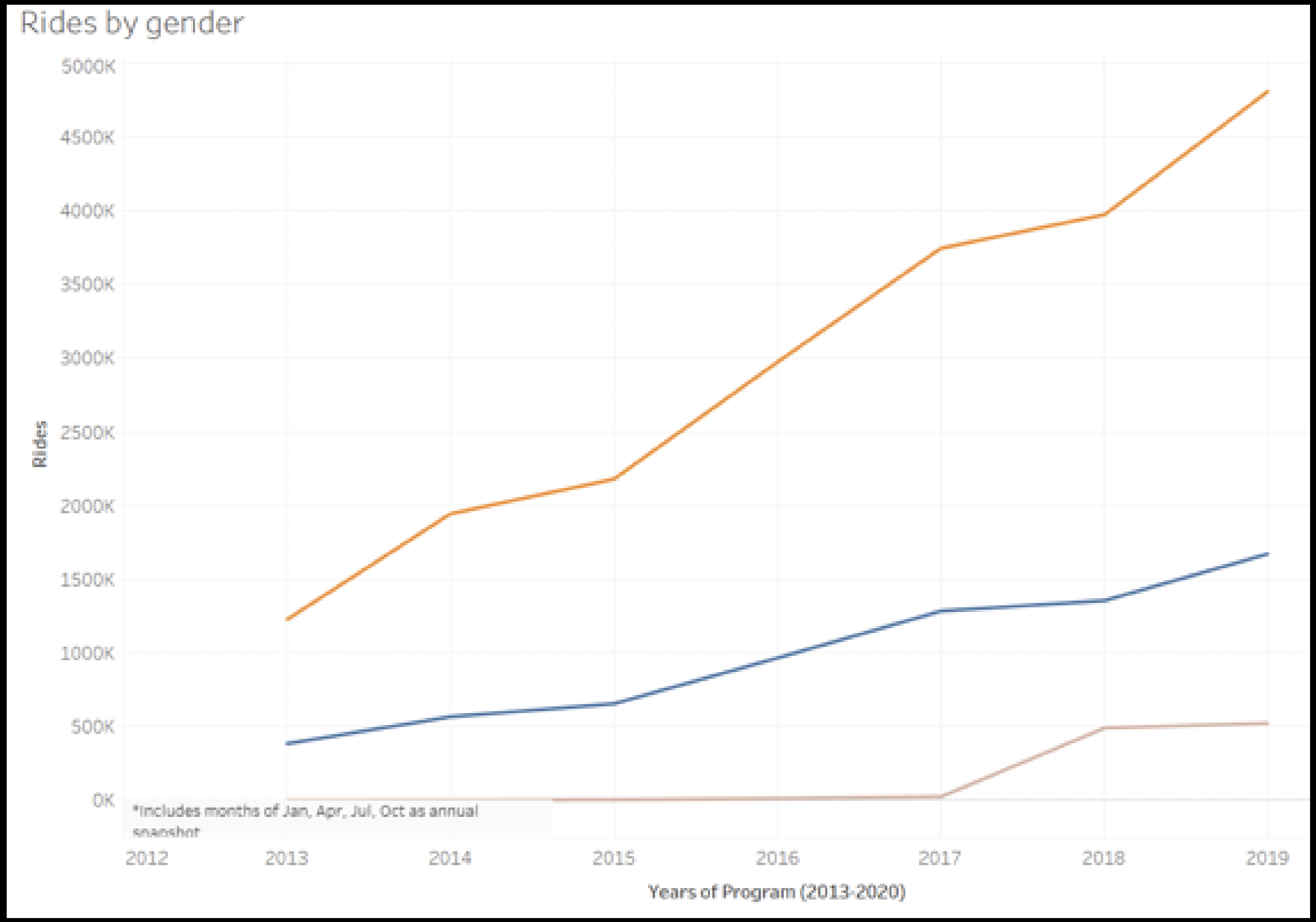
SOME FACTS

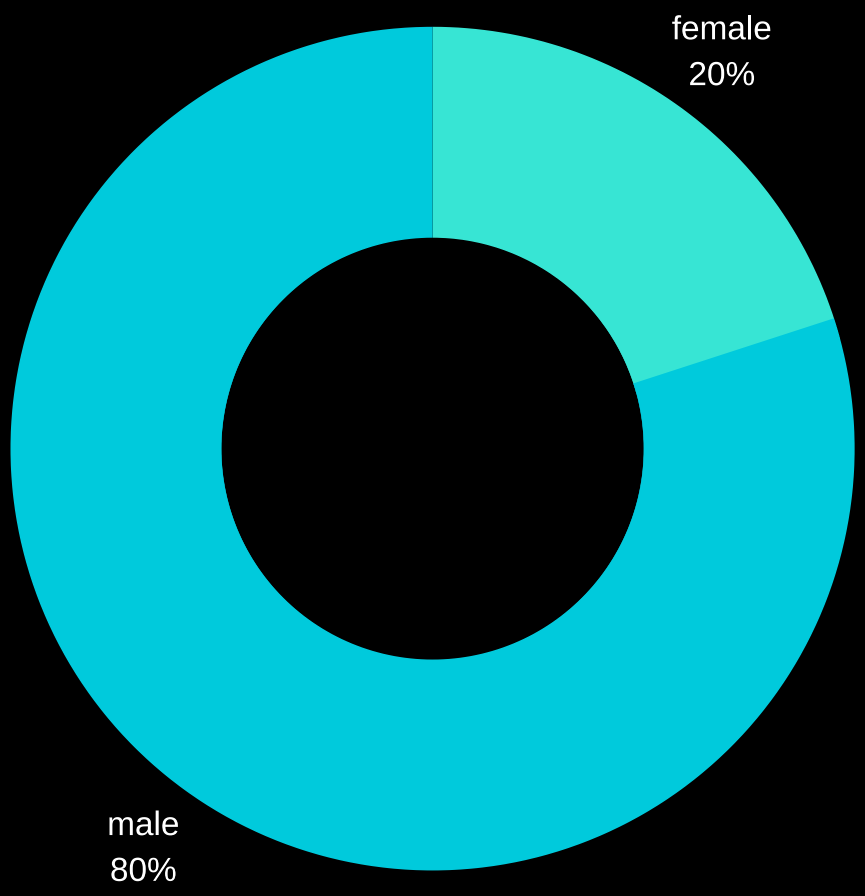
- According to recent studies, over half of females are the top financial decision maker of the household.
- Females make up more than half - 52% - of NYC's population.



SOME SAD FACTS

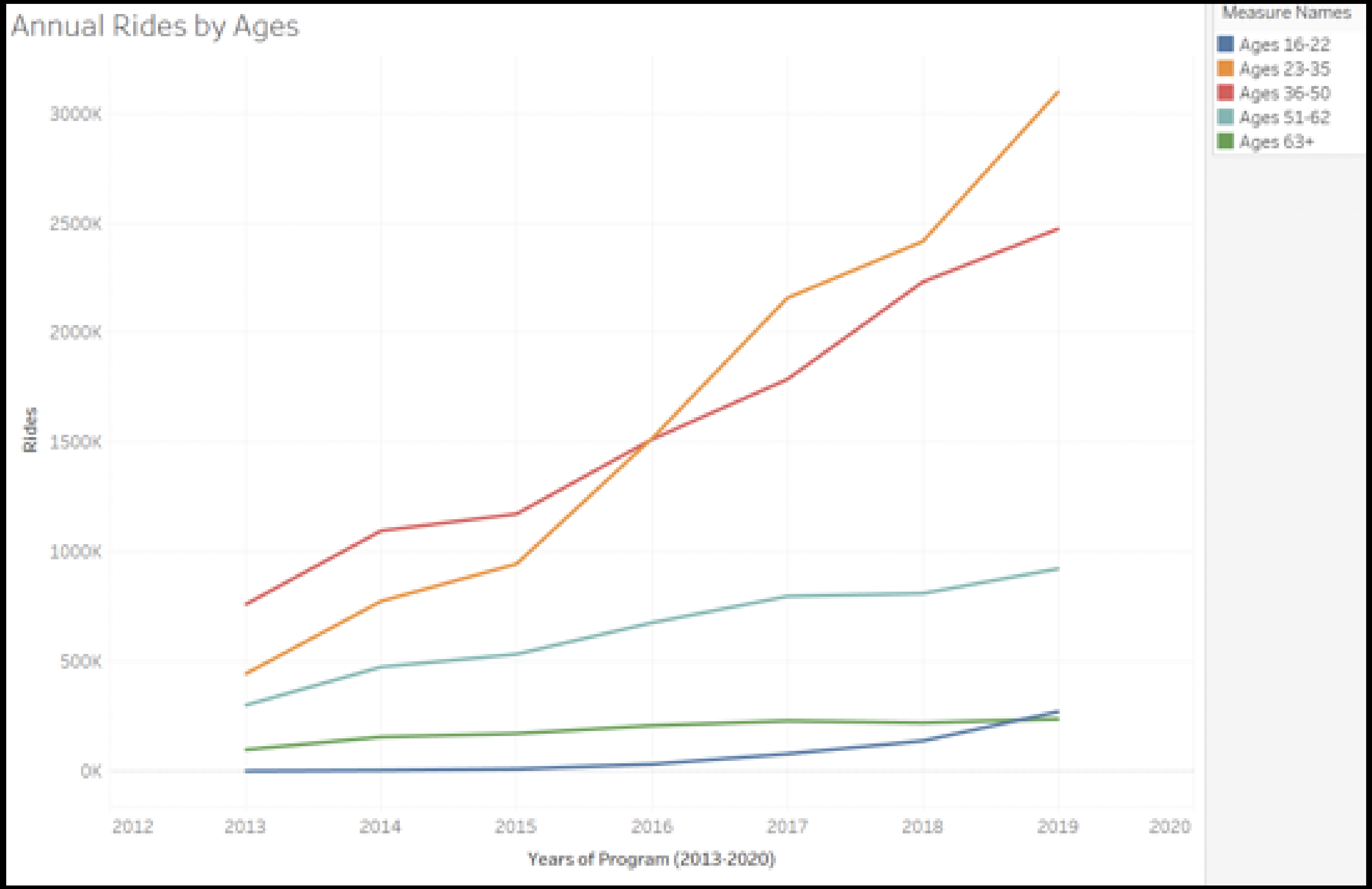
- According to our analysis females:male riders ratio is at 1:4 !
- The gap continues to increase year after years...





MORE SAD FACTS…

- Young riders from 16-22 years old are almost at the bottom place in terms of citibike usage.

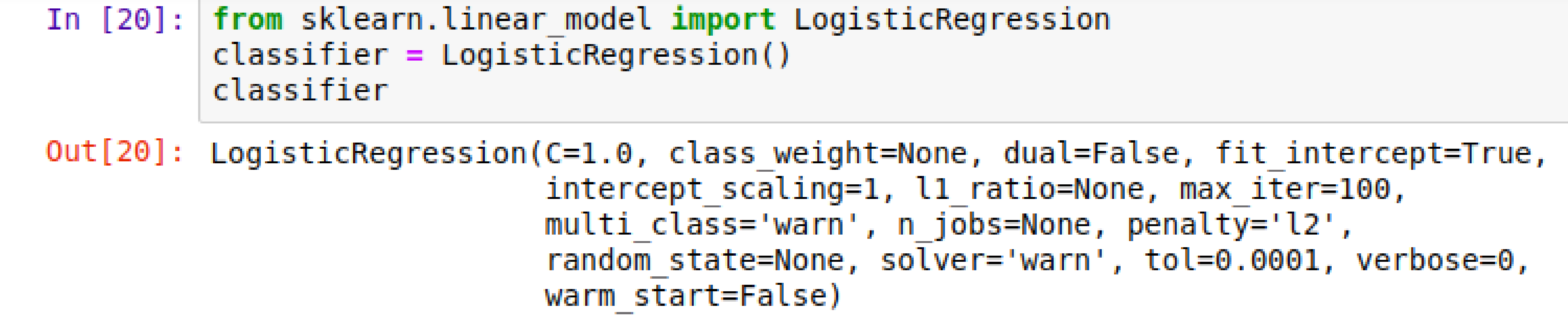


THE ANALYSIS

1. Use of Python / Pandas to import and clean CSV data.
2. Logistic regression:
   - Our model can predict the rider's gender given the start & end station, age and duration.
   - Random forest machine learning to determine the most important factors in determining the gender of a riders
3. Use of selenium and web scraping to get schools and yoga studios in NYC.
4. Tableau for data visualization, determine popularity of individual citibike stations, % of female to male riders, and to overlay maps of citibike stations with schools and yoga studios

THE PROCESS

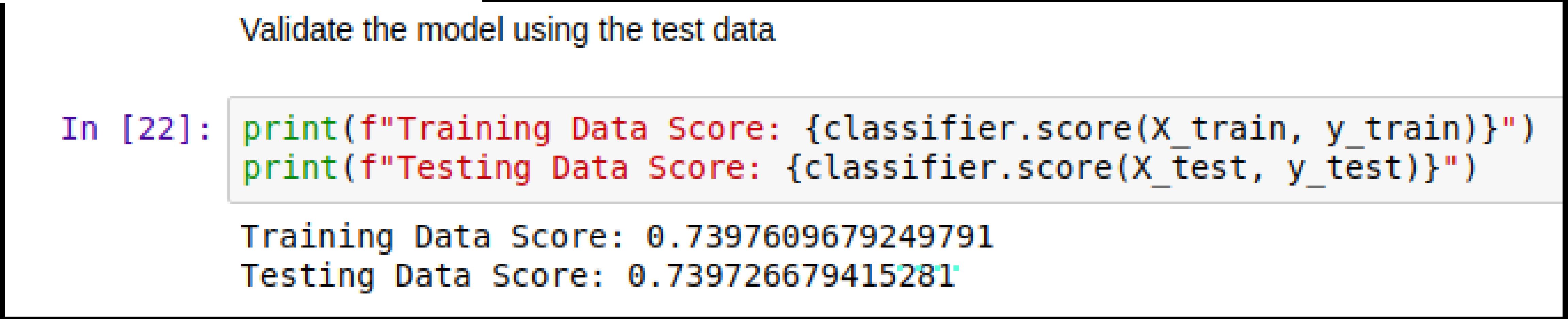
- In order to determine the weights of the factors that most greatly influence gender.
  - We used a binary logistic regression model from scikit-learn on a dataset of more than 1.4 million records from the Citibike data catalogue.
  - We used a standard 80% train / 20% test fit for our model.



```
In [20]: from sklearn.linear_model import LogisticRegression
         classifier = LogisticRegression()
         classifier

Out[20]: LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                            intercept_scaling=1, l1_ratio=None, max_iter=100,
                            multi_class='warn', n_jobs=None, penalty='l2',
                            random_state=None, solver='warn', tol=0.0001, verbose=0,
                            warm_start=False)
```

- Python,Scikit-learn



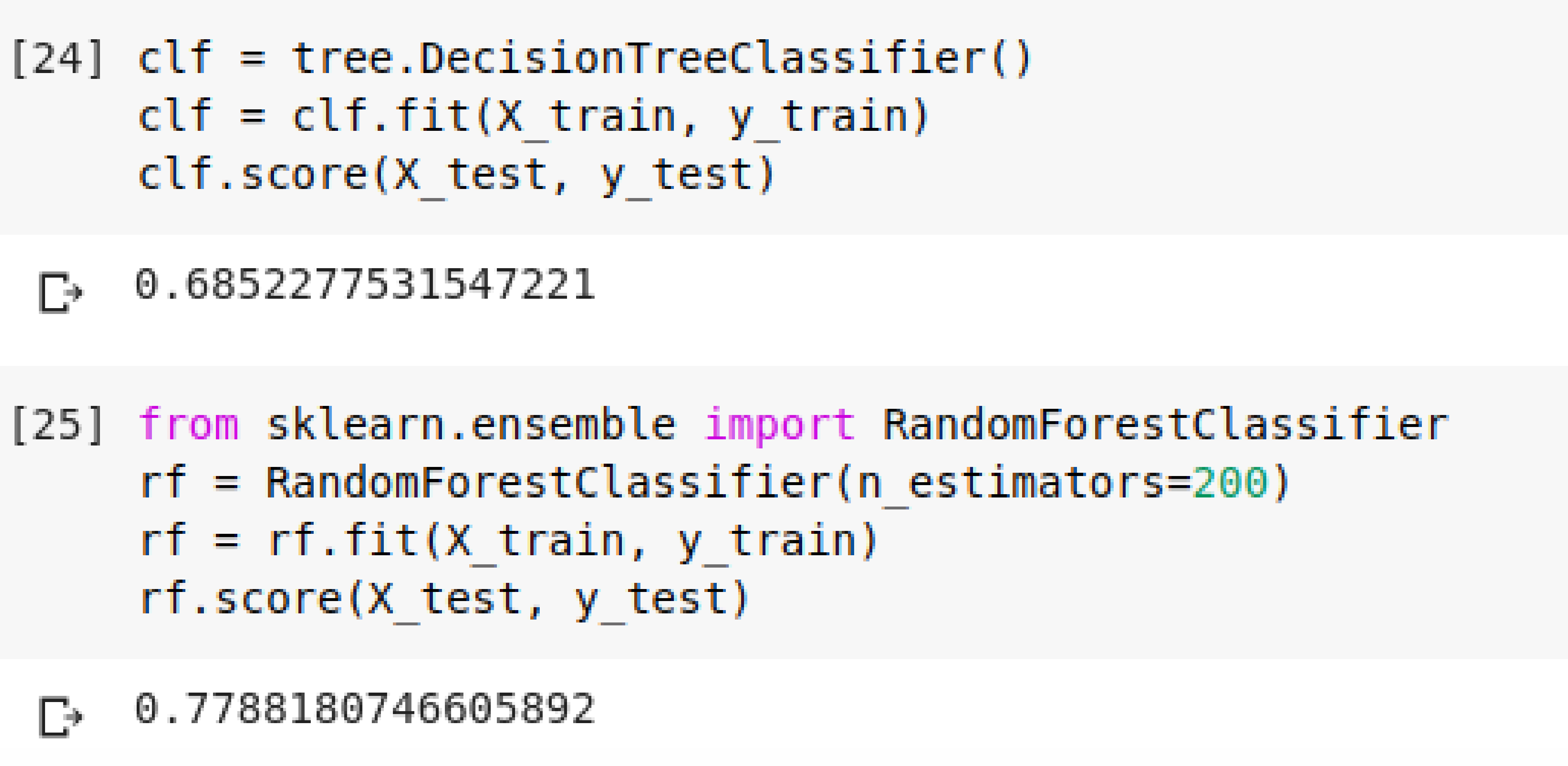
Validate the model using the test data

```
In [22]: print(f"Training Data Score: {classifier.score(X_train, y_train)}")
         print(f"Testing Data Score: {classifier.score(X_test, y_test)}")

         Training Data Score: 0.7397609679249791
         Testing Data Score: 0.739726679415281
```

THE PROCESS

- And random forests machine learning to determine the most important factors for 17-25 riders.



```
[24] clf = tree.DecisionTreeClassifier()
     clf = clf.fit(X_train, y_train)
     clf.score(X_test, y_test)
```

```
0.6852277531547221
```

```
[25] from sklearn.ensemble import RandomForestClassifier
     rf = RandomForestClassifier(n_estimators=200)
     rf = rf.fit(X_train, y_train)
     rf.score(X_test, y_test)
```

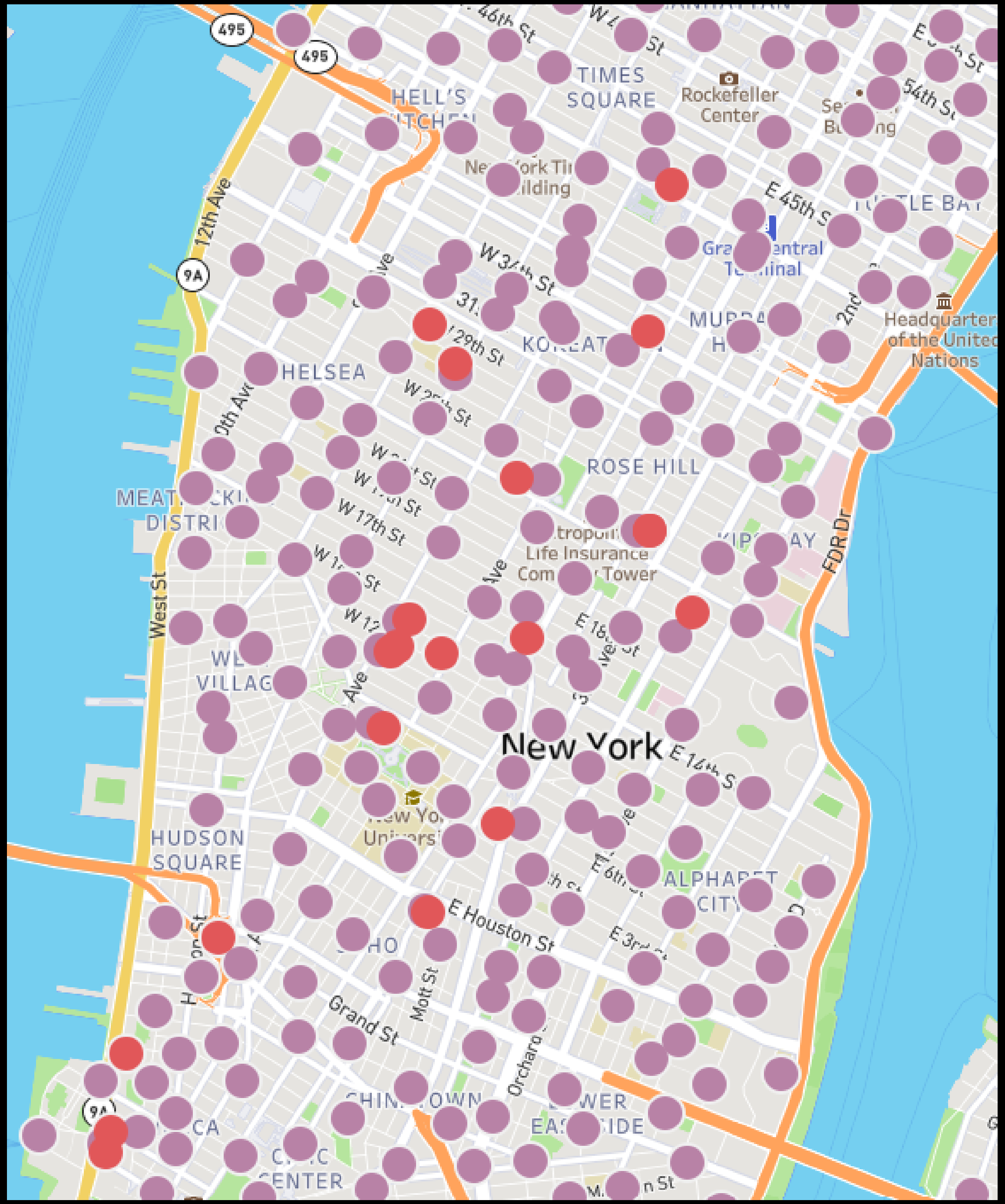
```
0.7788180746605892
```

THE PROCESS

- We calculated age of riders using the provided birth year and categorized them into age-groups

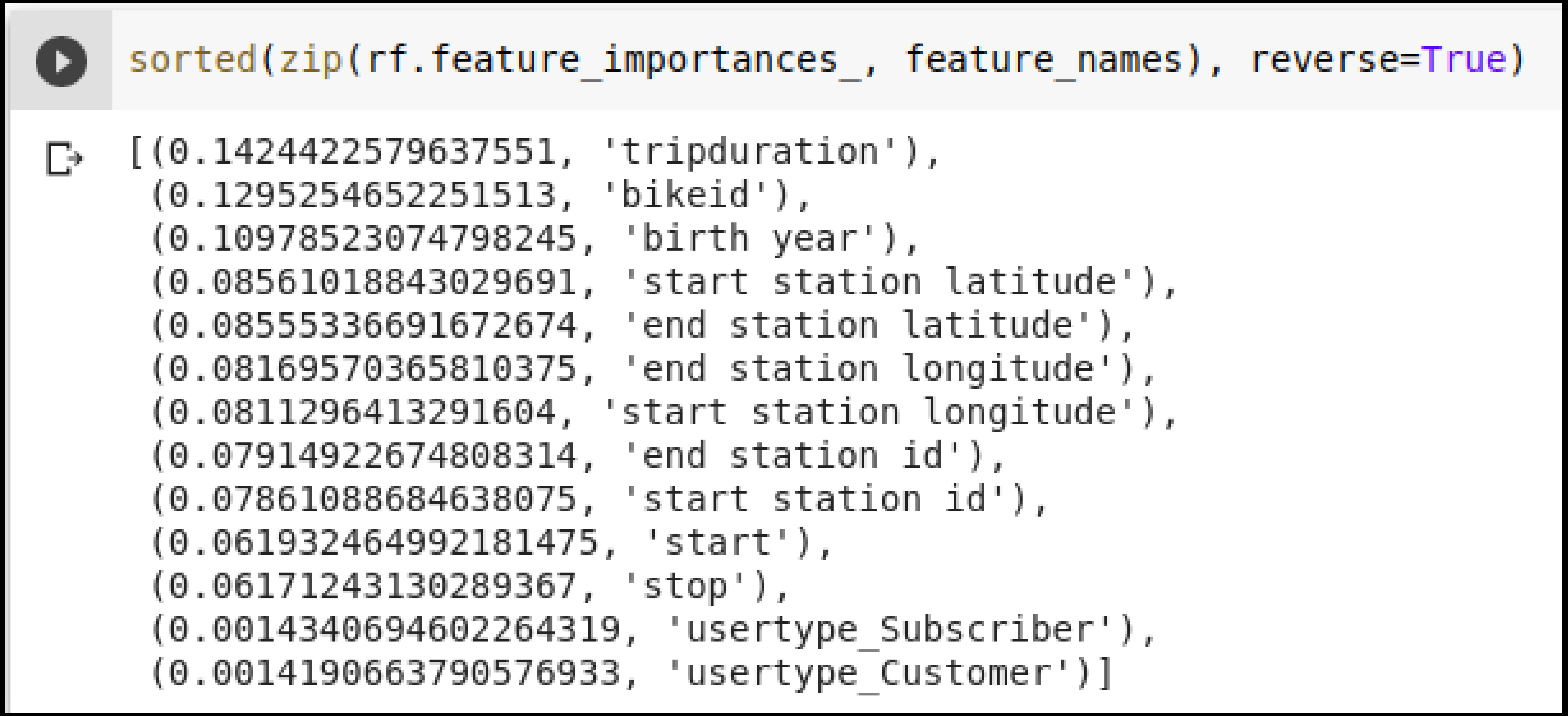
  (e.i., 0-16 (usertype=customer), 16-24, 25-50, 50-74, 74+).

  - We used SKlearn's random forest model again to determine the most important factors determining ridership for the age-group 16-24.
  - Mapping high schools, colleges and universities and selecting the top 5-10 locations where we would target campaigns toward this demographic.



SOME FINDINGS!… (GENDER)

- The most important features to determine **gender** are "trip-duration" and "bike id".

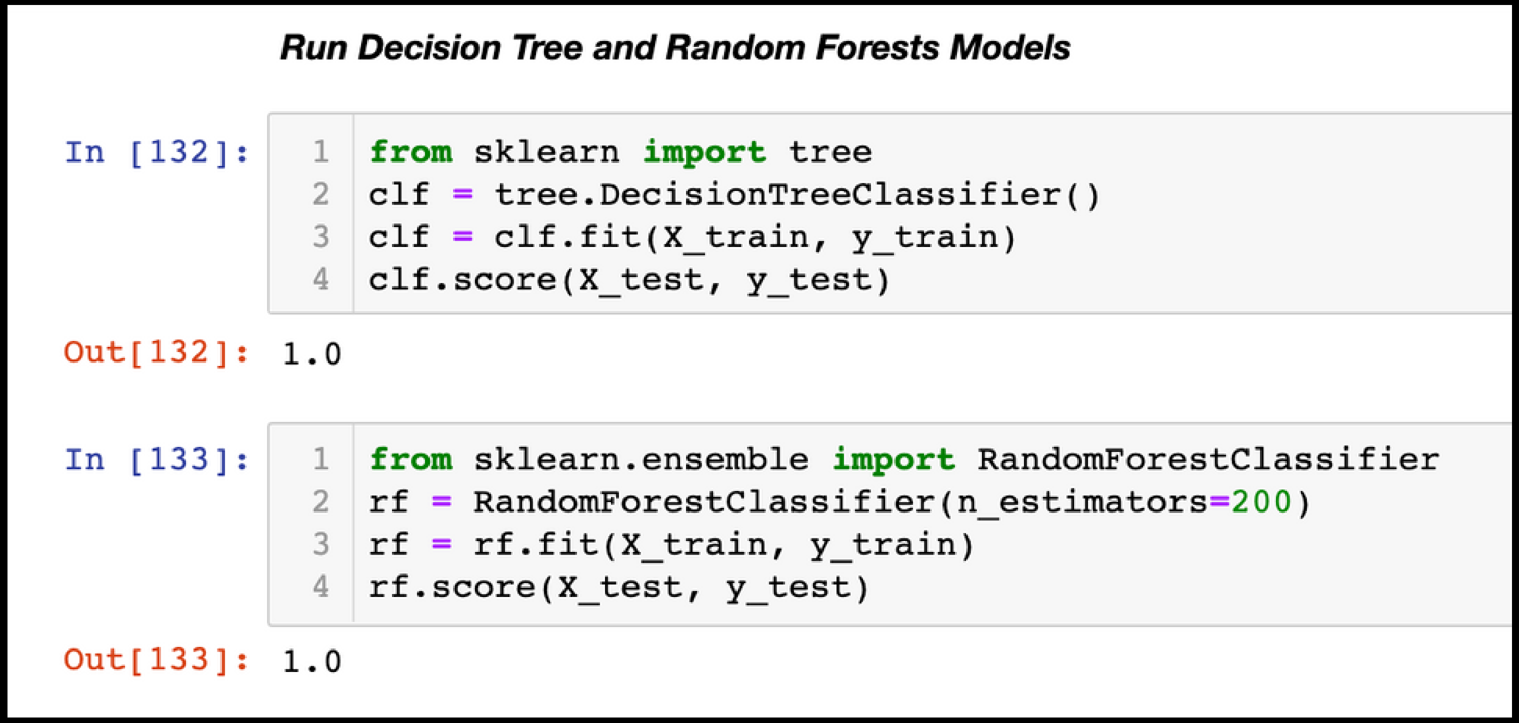


```
sorted(zip(rf.feature_importances_, feature_names), reverse=True)
```

```
[(0.1424422579637551, 'tripduration'),
 (0.1295254652251513, 'bikeid'),
 (0.10978523074798245, 'birth year'),
 (0.08561018843029691, 'start station latitude'),
 (0.08555336691672674, 'end station latitude'),
 (0.08169570365810375, 'end station longitude'),
 (0.0811296413291604, 'start station longitude'),
 (0.07914922674808314, 'end station id'),
 (0.07861088684638075, 'start station id'),
 (0.061932464992181475, 'start'),
 (0.06171243130289367, 'stop'),
 (0.0014340694602264319, 'usertype_Subscriber'),
 (0.0014190663790576933, 'usertype_Customer')]
```

SOME FINDINGS!... (AGE)

While the model was able to predict pretty accurately, the two most important factors seem to be stop_time (when the user ends the ride) and birth-year (which is a proxy for age-group). Whoops, lesson learned.. we should have dropped the birth year column to eliminate redundancies.



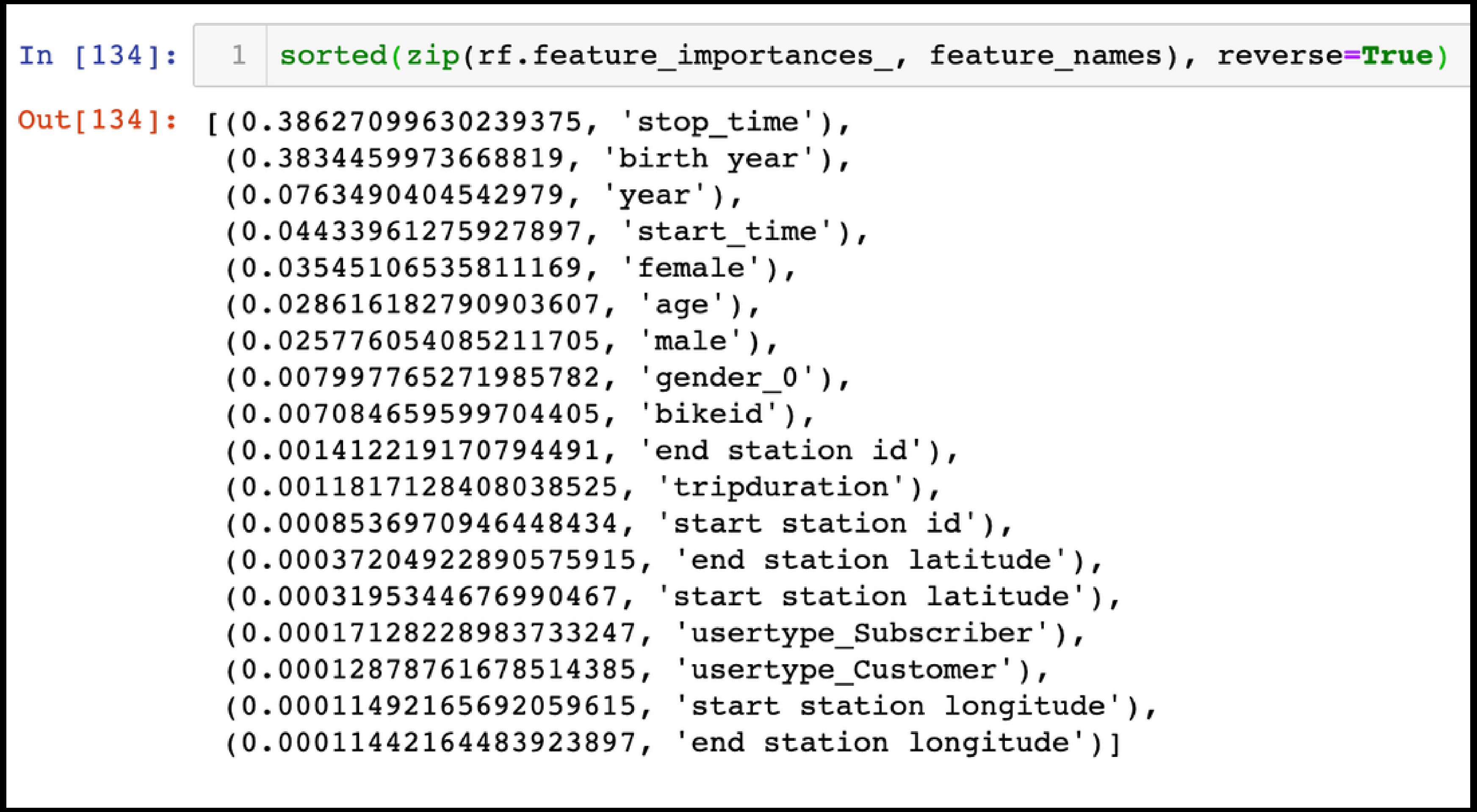
Run Decision Tree and Random Forests Models

```
In [132]:  from sklearn import tree
           clf = tree.DecisionTreeClassifier()
           clf = clf.fit(X_train, y_train)
           clf.score(X_test, y_test)

Out[132]:  1.0

In [133]:  from sklearn.ensemble import RandomForestClassifier
           rf = RandomForestClassifier(n_estimators=200)
           rf = rf.fit(X_train, y_train)
           rf.score(X_test, y_test)

Out[133]:  1.0
```



```
In [134]:  sorted(zip(rf.feature_importances_, feature_names), reverse=True)

Out[134]: [(0.38627099630239375, 'stop_time'),
           (0.3834459973668819, 'birth year'),
           (0.0763490404542979, 'year'),
           (0.04433961275927897, 'start_time'),
           (0.03545106535811169, 'female'),
           (0.028616182790903607, 'age'),
           (0.025776054085211705, 'male'),
           (0.007997765271985782, 'gender_0'),
           (0.007084659599704405, 'bikeid'),
           (0.001412219170794491, 'end station id'),
           (0.0011817128408038525, 'tripduration'),
           (0.0008536970946448434, 'start station id'),
           (0.00037204922890575915, 'end station latitude'),
           (0.0003195344676990467, 'start station latitude'),
           (0.00017128228983733247, 'usertype_Subscriber'),
           (0.00012878761678514385, 'usertype_Customer'),
           (0.00011492165692059615, 'start station longitude'),
           (0.00011442164483923897, 'end station longitude')]
```

All other factors (like year or ride start_time) have a much smaller influence.

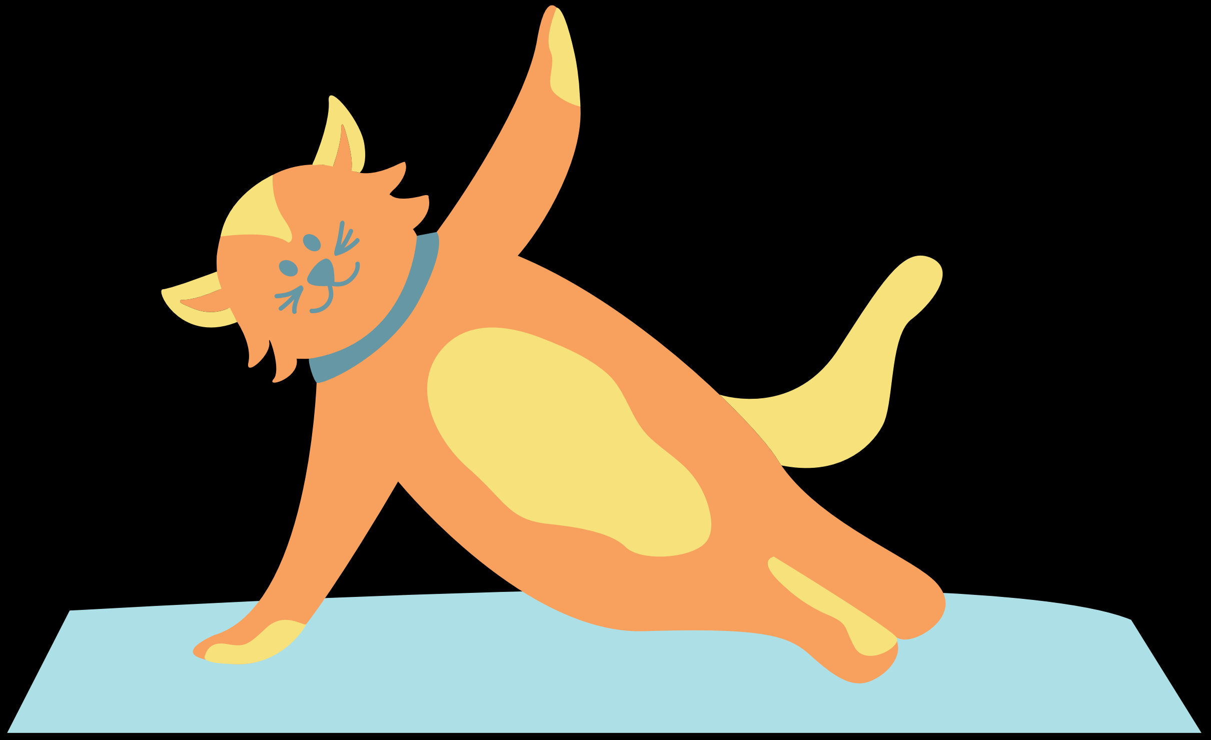
ACTIONS BASED ON OUR ANALYSIS:

- Deploy a campaign with promotions around:
  - yoga studios,
  - female-centric co-working spaces,
  - women professional groups,
  - “mompreneurs”
  - colleges / universities in NYC.

- Introduction of “design your own bike wrap” competitions
- Citibike “meetup rides” for female professional groups
- Video-testimonials from prominent female users

YOU MAY ASK, WHY YOGA?

- 72% of all yoga practitioners are female
- Yoga complements other forms of exercise: 79% of yogis also engage in exercise including running, group sports, weight lifting and cycling.
- The most active yoga age group is 30-49 year-olds, which aligns with the most frequent ages of Citibike riders.



DEEP DIVE INTO OUR OTHER AREAS OF FOCUS:

- The Wing: Co-work spaces for women with three locations in NYC: Williamsburg, Upper West Side, Bryant Park
- Women Get Paid: A private online network where thousands of women from around the world share advice, resources, and job opportunities.
- The Camaraderie: A women’s professional group founded by 25 year-old Jane Taylor that provides an open space for women free of pressures and pretenses, among hustle and bustle of New York City.

QUANTITATIVE AND MEASURABLE GOALS

- Increase female riders by 45% over the course of the next year.
- Decrease the existing gap of growth on female riders with male riders by 20% (currently at 3%).
- Increase 16-22 riders by 30% and have a steady growth of 10% each year.