**SLIDE 1**

Hello everyone, I would like to start this presentation by introduction, my name Sunaina,

So, my project is about….

* Clusterization techniques dataset - cab company – NYC

**SLIDE 2**

* Goals - identify
  + demand hotspots of passengers
  + time behaviors of the clients (rush hours of pick-up passengers)
  + perform a customer segmentation
    - insights brings benefits, such as:
      * Allocate strategic resources
        + NYC good locations for stand of taxis
        + drivers shift distribution
        + to understand the client and provide a better service
        + increasing profitability
        + Improving
        + operational efficiency

**SLIDE 3**

So, our first step of this project was to merge the datasets of 3 different months of the year 2016 into one dataset.

* +34 million observations
* Stratified sample
  + 10% dataset/easier to handle
  + categorical/type of payment class balance
* EDA
  + No missing values, no duplicated observations
  + Several outliers – almost all variables
* OUTLIERS – CLEANING – IMPORTANT POINT
  + MIN – MAX range the goal of this project is to track hotspots to pick up passages
  + Decided to restrict the cleaning because
    - One of the variables is distance, in practical terms, if you go to an airport using a cab, it is different than to go to a friend’s house
    - unrealistic values would be removed
      * one of the observations in the dataset were located using the coordinates of New Haven, New Haven is a city close to New York, but as the coordinates of this observation were out of the New York City coordinates, these observations got removed.

For the next step, we evaluated different methods to check if this dataset resulting dataset is “clustable”, which will be presented by….

**SLIDE 4**

* PCA
  + Dense concentration horizontal axis
  + not randomly distributed
  + good tendency for clustering
* HOPKINS
  + Almost 85%
* DISTANCE MATRIX
  + Blue means that observations are close to each other
  + Red means anomalies (it maybe be an airport for example, which is far from the city)

**SLIDE 5**

* “K” Winner – 2 methods
  + Elbow Method
    - Based on the graph k = 4
      * After k = 4, the lines significantly slow down
      * Indicates the elbow point
  + Silhouette Analysis
    - Peak = 2
    - 4 also high
* “K” Winner = 4