**Data Mining Project**

**Problems:**

1. **The average miles travelled is 10.4. A random sample of 40 customers is collected by using 95% confidence level, validate the hypothesis through at least 3 approaches.**

Here we want to examine if the average miles traveled by a passenger is 10.4 or not.

Null Hypotheses: Average miles traveled by an Uber customer is 10.4 miles.

Alternate Hypotheses: Average miles traveled by an Uber customer is not 10.4miles.

H0 = µ=10.4

HA = µ ≠ 10.4

CODE:

PB1=read.csv("DM\_PRJ\_Q1.csv", header=TRUE,sep=",")

Summary(PB1)

MILES=PB1$MILES

library(BSDA)

z.test(MILES,NULL,alternative="two.sided",mu=10.4,sigma.x=sd(MILES),conf.level=0.95)

One-sample z-Test

data: MILES

z = -0.73173, p-value = 0.4643

alternative hypothesis: true mean is not equal to 10.4

95 percent confidence interval:

5.516755 12.628245

sample estimates:

mean of x

9.0725

> summary(MILES)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.800 2.400 5.650 9.072 10.500 63.700

> SD(MILES)

Error in SD(MILES) : could not find function "SD"

> save=sd(MILES)

> save

[1] 11.47394

histogram(MILES,xlab="MILES",ylab="CUSTOMER'S DRIVE")

**Manual calculations:**

**By Test Statistic Approach:-**

# Z= X- µ / σ/

X = 9.072

µ= 10.4

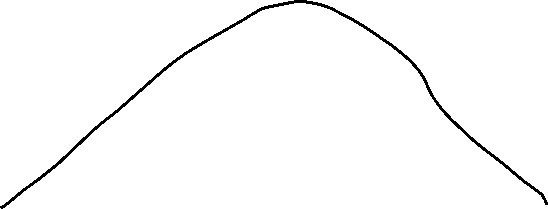
σ= 11.473

= 6.32

Z = (9.072-10.4)/ (11.473/6.32)

= -1.38 / 1.815 = -0.7603

Z=-0.76



-1.96 +1.96

From the bell curve above, we see that Zc falls in acceptance region. So, we will accept null hypothesis and reject alternative hypothesis.

We will say that the average miles traveled by a uber customer is 10.4miles.

**BY CONFIDENCE INTERVAL METHOD:**

X = 9.072

σ = 11.473

n =40

√40 = 6.32

confidence = 95%

X ± Zα \* (σ/√N))

=X + Zα \* (σ/√N)) = 9.072 + 1.96\*(11.473/6.32) or 9.072 - 1.96\*(11.473/6.32)

= 9.072 + 1.96 (1.815) or 9.072 – 1.96(1.815)

=9.072 + 3.557 or 9.072 – 3.557

=12.629 or 5.515

Interval = [12.629, 5.515]

5.515 µ 12.629

µ = 10.4 is present in the interval [12.629, 5.515].

We do not have enough evidence to reject Null hypothesis and hence we accept Null hypothesis under a confidence of 95%.

We will say that the average miles traveled by a uber customer is 10.4miles.

**By P-Value Approach:**

https://www.graphpad.com/quickcalcs/pValue2/

P-Value 0.4473

Comparing p and α,

α = 0.05

P > α

We do not have enough evidence to reject Null hypothesis and hence we accept Null hypothesis under a confidence of 95%.

We will say that the average miles traveled by a uber customer is 10.4miles.

**2.) For Business user’s compare average fare is greater than average fare for Personal user’s.**

**It is two sample independent hypothesis testing. Also, it is a two-tailed test We will mention null and alternate hypothesis. Since sample size is greater than 30 so we will use z-distribution hypothesis testing to find whether mean fare of Business users is greater than mean fare of personal users. It is assumed that mean fare for both users is same, but we do not think so.**

Fare of Customers using Uber for Business use

Sample size:23

Fare of Customers using Uber for Personal use

Sample size:17

**CODE:**

Business=read.csv("Business\_Fare.csv", header=TRUE,sep=",")

summary(Business)

Fare = Business $ Fare

summary(Fare)

> Fare\_sd=sd(Fare)

> Fare\_sd

[1] 11.52416

library(psych)

describe(Fare)

**Personal Fare:**

Personal=read.csv("Personal\_Fare.csv", header=TRUE,sep=",")

summary(Personal)

PFare = Personal $ Fare

summary(PFare)

> PFare\_sd=sd(PFare)

> PFare\_sd

[1] 11.83216

library(psych)

describe(PFare)

Null Hypothesis: Average fare for business users is same as the average fare for personal users.

Alternate Hypothesis: Average fare for business users is not same as the average fare for personal users.

µ1 = Average fare for business users.

µ2= Average fare for personal users.

H0 = µ1 = µ2

HA = µ1 ≠ µ2

i.e

H0 = µ1 - µ2 = 0 i.e µd=0

HA = µ1 - µ2 ≠ 0 i.e µd ≠ 0

library(BSDA)

> t.test(Fare,PFare,alternative="two.sided",mu=0,paired=F,var.equal=T,conf.level=0.95)

Two Sample t-test

data: Fare and PFare

t = -0.93307, df = 38, p-value = 0.3567

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-11.024715 4.068193

sample estimates:

mean of x mean of y

20.52174 24.00000

P value > α i.e 0.35 > 0.05

Hence, we accept null hypothesis

**Manual Calculations:**

**X1 = 20.52**

**X2 = 24**

µd = 0

σ1 = 11.52

σ2 =11.83

n1 = 23

n2= 17

√17 =4.12

√23 = 4.79

Df = 23+17-2 = 38

Critical value t =?

Sd =(n1-1) Sd^2 + (n2-1) Sd2^2 / (n1-1) + (n2-1)

= 22 (11.52) ^2 + (16) (11.83)^2 / 38

= 2919.62 + 2239.18 / 38

Sd = 135.75

T =(X1-X2) –(µ1-µ2) /

= (20.52 - 24) – (0) /

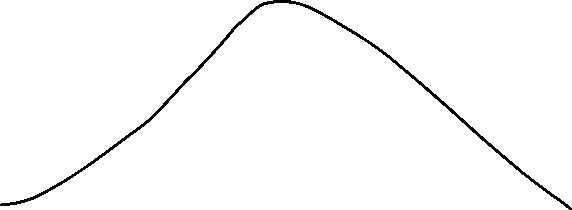
= -3.48 /

= -3.48 /

= -3.48 /43.312

= -0.0803

Z=7.404



-2.024 +2.024

**3.) Using classification find which age group prefers more cabs(categories, purpose, gender and riders age)**

**It is classification problem. We will use different classification algorithms like KNN and Naive Bayes to find which age group people prefer more cabs and will compare results for all the algorithms. We will take different algorithms and compare algorithm gives us best results. Here we will consider label as Age and the values under age will be Y (Young), M (Middle Age), O (Old Age).**

# -\*- coding: utf-8 -\*-

"""

Created on Thu Apr 19 00:31:27 2018

@author: neera

"""

#import libraries

import numpy as np

import matplotlib.pyplot as plt

import pandas as pd

# importing our csv dataset

mydata=pd.read\_csv('classification\_problem.csv')

X=mydata.iloc[:,[0,1,2]].values

y=mydata.iloc[:,[3]].values

# splitting data into training and testing data

from sklearn.cross\_validation import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X,y, test\_size=0.25, random\_state=0)

# Feature scaling

from sklearn.preprocessing import StandardScaler

sc=StandardScaler()

X\_train=sc.fit\_transform(X\_train)

X\_test=sc.transform(X\_test)

#Classification algorithm area

from sklearn.naive\_bayes import GaussianNB

classifier=GaussianNB()

classifier.fit(X\_train,y\_train)

# get output of test results

y\_pred=classifier.predict(X\_test)

#create confusion matrix

from sklearn.metrics import confusion\_matrix

cm=confusion\_matrix(y\_test,y\_pred)

# visualize the training data

from matplotlib.colors import ListedColormap

X\_set,y\_set =X\_train,y\_train

#for i,j in enumerate(np.unique(y\_set)):

#plt.scatter(X\_train[:,0],y\_train[:,1])

# plt.scatter(X\_set[y\_set==j,0],X\_set[y\_set==j,1],c=ListedColormap(('red','green'))(i),label=j)

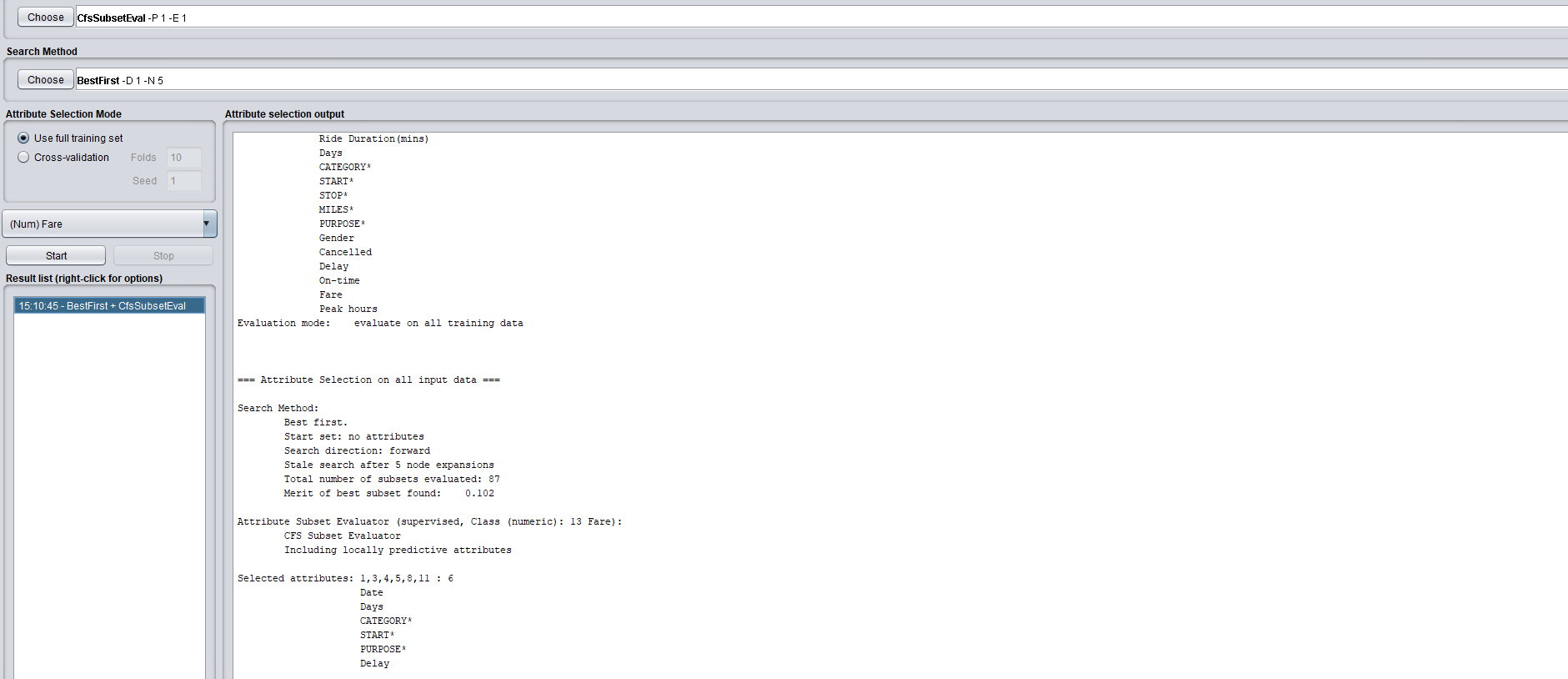
plt.scatter(X\_train[:,0],y\_test[:,1])

plt.title('Naive Bayes Visualization')

plt.show()

**4:** **Select the features which are used to estimate cab fare?**

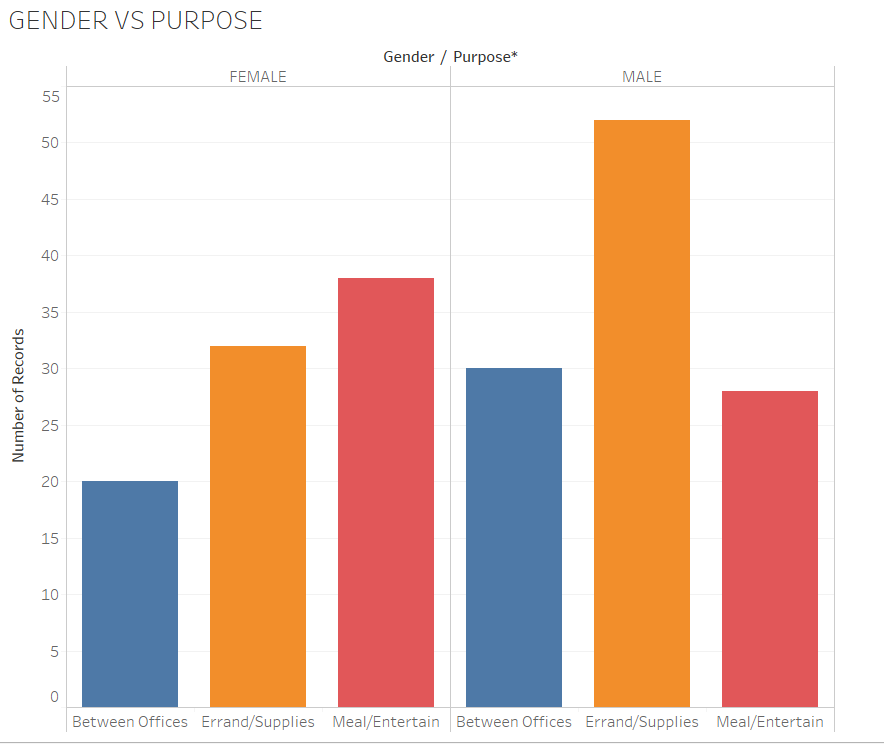
**Solution 4: It is features selection and feature reduction problem where we will eliminate the features which are not useful for finding the cab fare.For example: we will remove the feature gender to calculate cab fares as fares is not dependent on the person who is travelling.**

****

**5) Perform a cross tabulation of two gender variables versus three purpose variables?**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| **GENDER** | **PURPOSE=Between Offices** | **PURPOSE=Errand/Supplies** | **PURPOSE=Meal/Entertain** |
| MALE | 30 | 52 | 28 |
| FEMALE | 20 | 32 | 38 |

|  |  |  |
| --- | --- | --- |
| **PURPOSE\*** | **GENDER** | **Count of PURPOSE\*** |
| **Between Offices** | FEMALE | 20 |
| **Between Offices** | MALE | 30 |
| **Errand/Supplies** | FEMALE | 32 |
| **Errand/Supplies** | MALE | 52 |
| **Meal/Entertain** | FEMALE | 38 |
| **Meal/Entertain** | MALE | 28 |

****

From the graphs and data, we can say,

Males prefer Uber maximum for using it between offices.

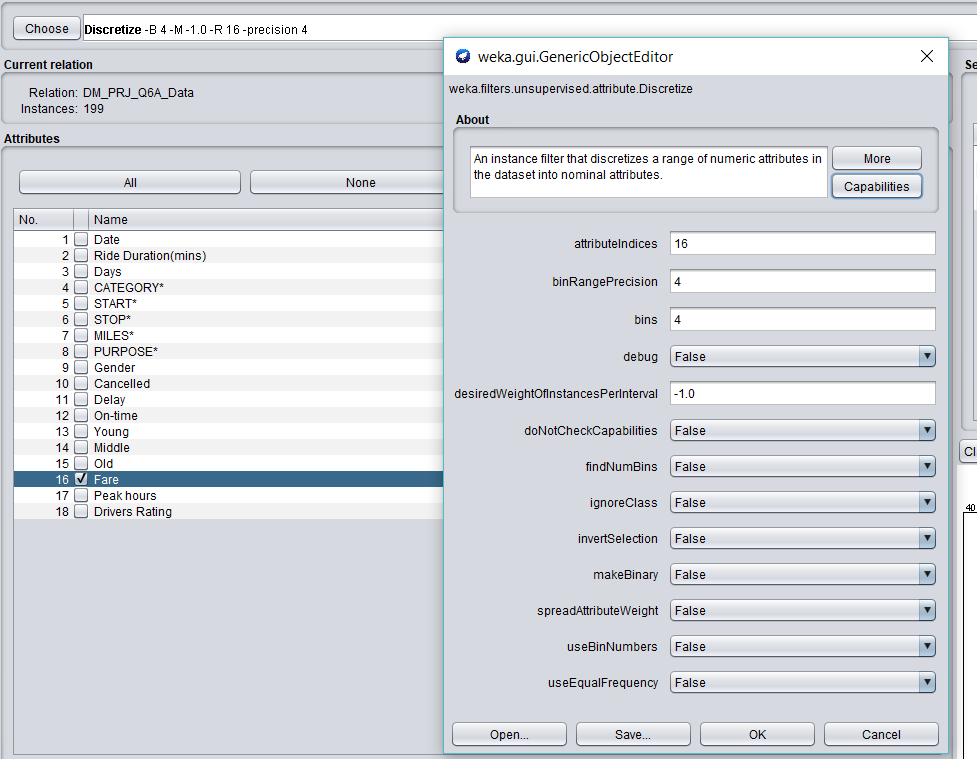
Males prefer Uber maximum for using it In Errand/Supplies.

Females prefer Uber maximum for using It for Meal/Entertainment.

**6) a) Using Smoothing by bin means on Fare attribute. b) Discretize the age attribute by following categories: Young: 1-30, Middle: 31-60, Old: 61-90**

**Solution6: a) In this we will do smoothing by bin means method on fare attribute. We will divide the date in the bin of 4 and then replace by the mean value. b) In this we discretize the age attribute to convert numerical data to nominal data by following categories: Young: 1-30; Middle: 31-60; Old: 61-90**

**a.)**

****

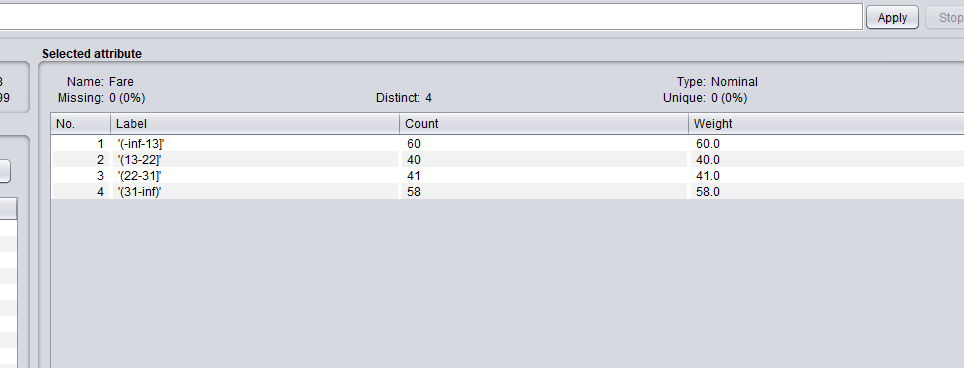
**4-Bins:**

0-13 – Bin 1

13-22 – Bin 2

22-31 – Bin 3

31-40 – Bin 4



**b) Discretize the age attribute by following categories: Young: 1-30, Middle: 31-60, Old: 61-90**

already done..

**7) Analyze the average ride duration.**

We shall follow below steps to build a regression model which best predicts the average ride duration.

1) Draw a scatter plot Y and X-variables.

2) Observe whether there is linear relationship or not.

3) Build a multiple linear regression model.

4) Parameter estimates.

5) Goodness of fit test.

6) Residual analysis.

7) Evaluation and prediction of best model.

**For Average ride duration, we require following independent variables:**

**Days**

**Start**

**Stop**

**Miles**

**Purpose**

**Peak hours**

**Category**

**Status**

**(Not applying model evaluation right now)**

AvgFare=read.csv("DM\_PRJ\_Q7\_DATA.csv", header=TRUE,sep=",")

summary(AvgFare)

Duration=AvgFare$Ride.Duration

Days=AvgFare$Days

Category=AvgFare$CATEGORY

Start=AvgFare$START

Stop=AvgFare$STOP

Miles=AvgFare$MILES

purpose=AvgFare$PURPOSE

cancelled=AvgFare$Cancelled

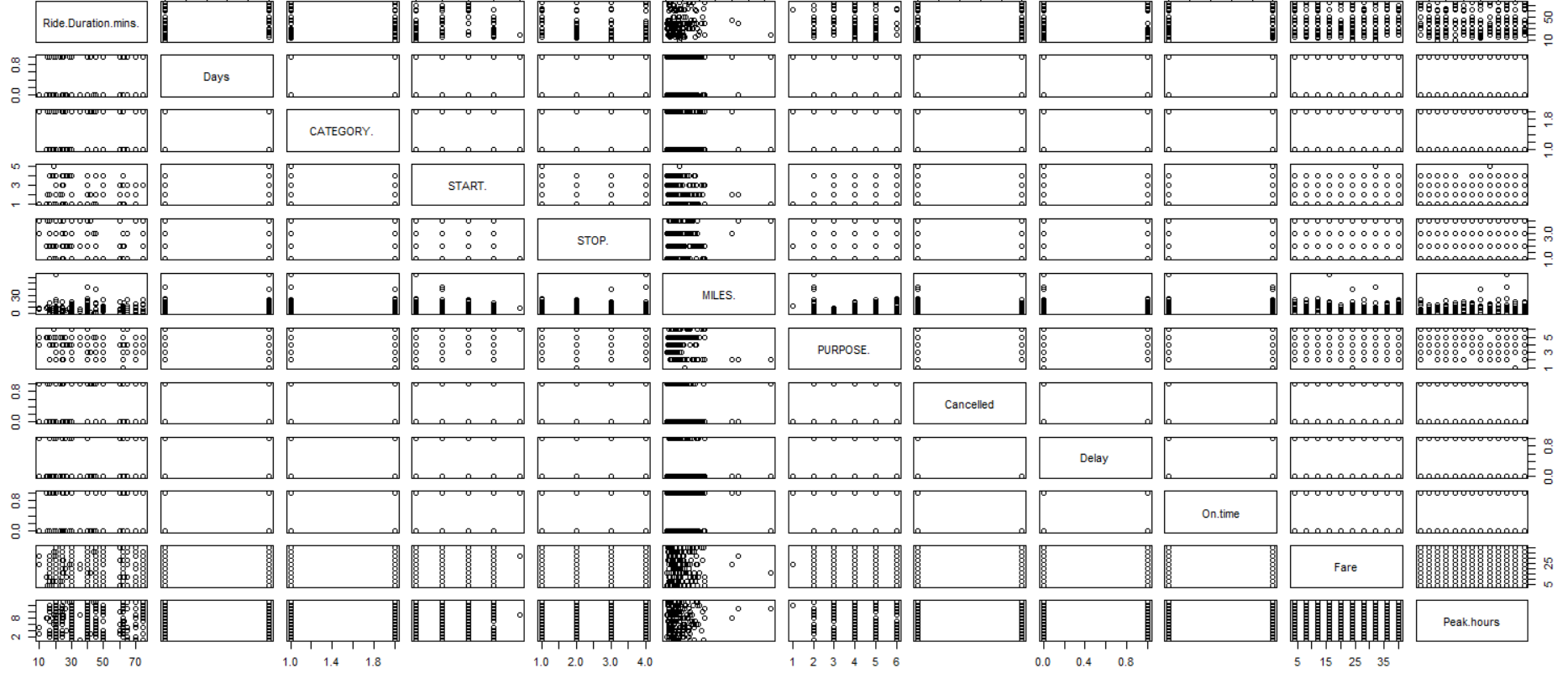
delay=AvgFare$Delay

ontime=AvgFare$On.time

fare=AvgFare$Fare

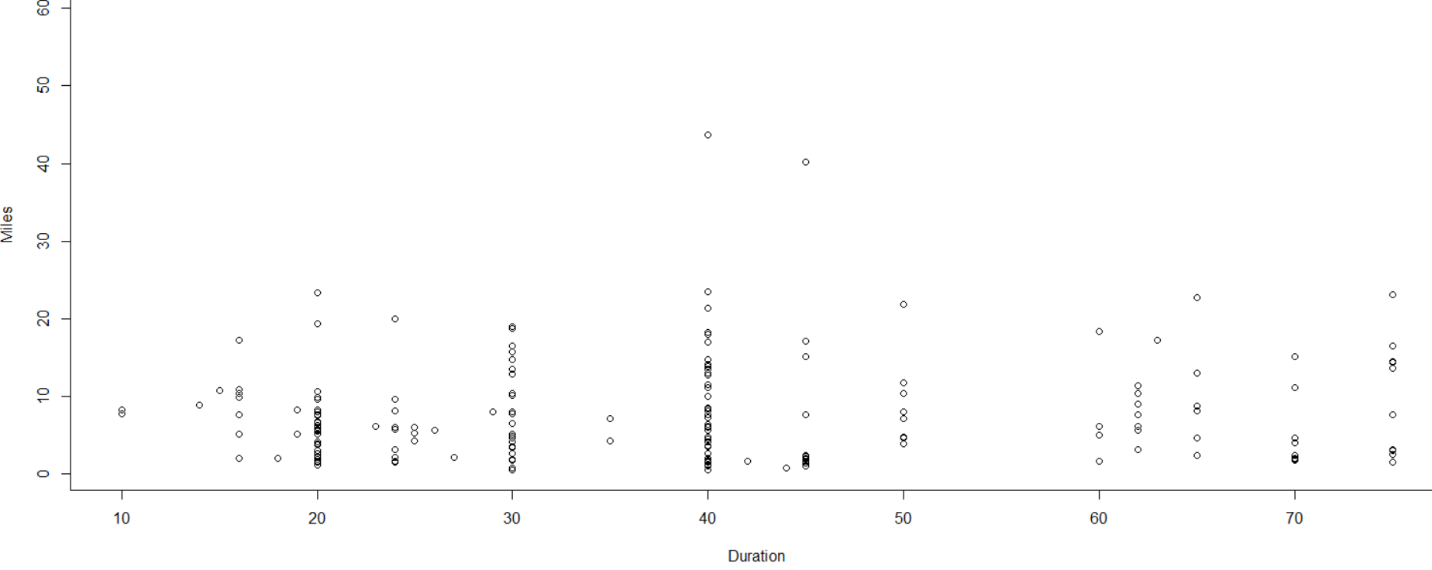
peakhours=AvgFare$Peak.hours

plot(AvgFare)

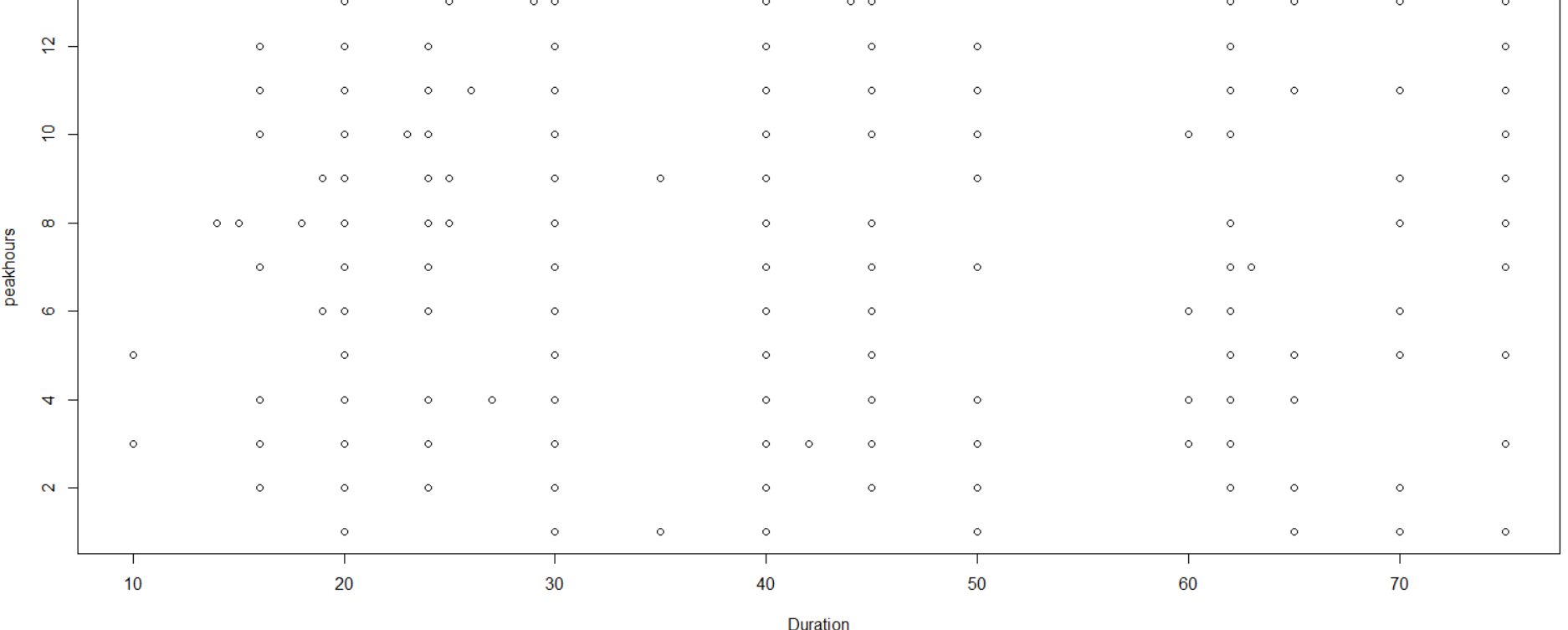


**We will examine the relationship between duration and miles,**

**Plot(Duration,Miles)**

****

**Plot(Duration,Peakhours)**

****

**Multiple linear regression model:**

**fit=lm(Duration~** **Days+** **Category+** **Start+Stop+** **Miles+** **purpose+** **cancelled+** **delay+** **ontime+** **fare+** **peakhours,data=** **AvgFare)**

> fit=lm(Duration~ Days+ Category+ Start+Stop+ Miles+ purpose+ cancelled+ delay+ ontime+ fare+ peakhours,data= AvgFare)

> summary(fit)

Call:

lm(formula = Duration ~ Days + Category + Start + Stop + Miles +

purpose + cancelled + delay + ontime + fare + peakhours,

data = AvgFare)

Residuals:

Min 1Q Median 3Q Max

-38.174 -11.535 -0.737 10.012 33.014

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 84.24060 24.36497 3.457 0.000692 \*\*\*

Days 0.47725 2.51778 0.190 0.849891

Categorypersonal -2.40791 2.48895 -0.967 0.334723

StartNorth 6.00206 3.42762 1.751 0.081767 .

StartSouth 18.03646 3.58669 5.029 1.26e-06 \*\*\*

StartWest -4.15659 3.35792 -1.238 0.217511

StartWest -18.26244 17.37196 -1.051 0.294658

StopNorth -2.27271 3.76799 -0.603 0.547219

StopSouth -4.60993 3.43153 -1.343 0.180963

StopWest 3.06153 3.91634 0.782 0.435480

Miles 0.09346 0.18378 0.509 0.611723

purposeCustomer Visit -39.49224 17.11162 -2.308 0.022230 \*

purposeErrand/Supplies -29.60107 17.04970 -1.736 0.084379 .

purposeMeal/Entertain -34.00106 16.78849 -2.025 0.044433 \*

purposeMeeting -34.55698 16.84937 -2.051 0.041836 \*

purposeTemporary Site -30.57646 17.07498 -1.791 0.075150 .

cancelled -9.64626 17.39434 -0.555 0.579935

delay -14.07350 17.14488 -0.821 0.412898

ontime -10.30398 17.13727 -0.601 0.548481

fare -0.05979 0.10339 -0.578 0.563850

peakhours 8:00PM 1.59204 5.97675 0.266 0.790282

peakhours10:00 AM -5.30195 5.78993 -0.916 0.361135

peakhours10:30 AM -2.19376 5.88095 -0.373 0.709600

peakhours6:00 PM -1.89165 6.06689 -0.312 0.755582

peakhours6:30 PM -3.67611 6.00261 -0.612 0.541095

peakhours7:00 PM -1.37518 5.81923 -0.236 0.813476

peakhours7:30 PM -3.63734 6.02312 -0.604 0.546732

peakhours8:00 AM -5.43753 6.02345 -0.903 0.367971

peakhours8:00 PM -1.39723 5.90485 -0.237 0.813238

peakhours8:30 AM 2.92047 6.14566 0.475 0.635259

peakhours8:30 PM -7.36384 6.04070 -1.219 0.224549

peakhours9:30 AM 3.22553 5.88113 0.548 0.584112

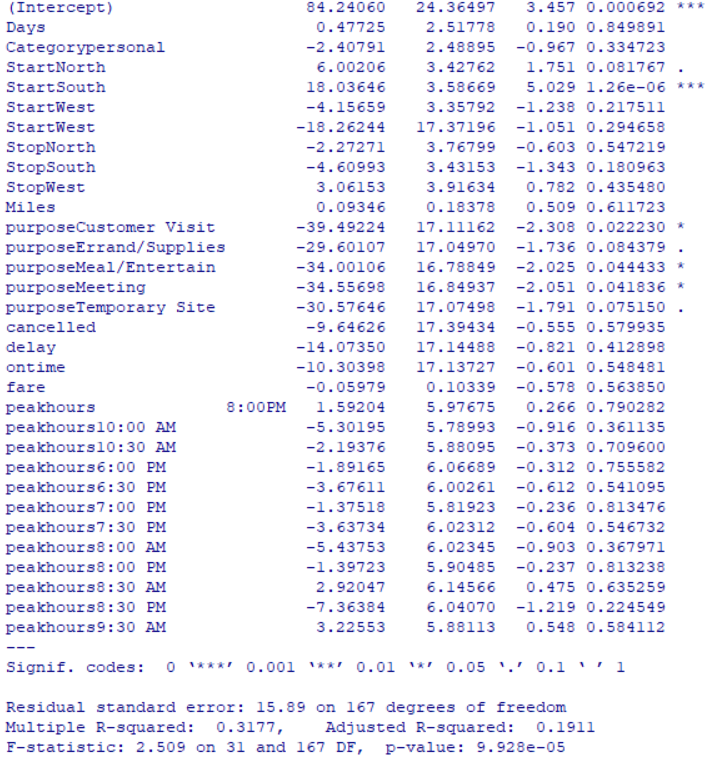
---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.89 on 167 degrees of freedom

Multiple R-squared: 0.3177, Adjusted R-squared: 0.1911

F-statistic: 2.509 on 31 and 167 DF, p-value: 9.928e-05



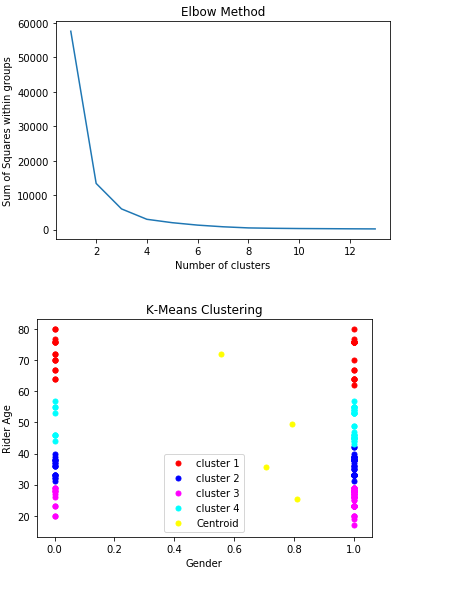
Removing parameters and building new model,

???????????????????????????????????????????

**8) Using clustering analyze which age group uses more cabs based on the category.(Age vs Gender)**

**Solution 8: It is clustering problem. We will use different clustering algorithms like partitioning and hierarchical we can interpret that which age group uses cabs for which and will analyze them to find the results.**

**USING PYTHON:**

****

**CODE:**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

from matplotlib.colors import ListedColormap

from sklearn.cluster import KMeans

# importing our csv dataset

mydata=pd.read\_csv('DM\_PR8.csv')

X=mydata.iloc[:,[6,13]].values

# Find out the best number of clusters

Array=[] # to store sum of squares within the groups

for i in range(1,14):

kmeans=KMeans(n\_clusters=i,init='k-means++',random\_state=0)

kmeans.fit(X)

Array.append(kmeans.inertia\_) # inertia --> Sum of squared distances of samples to their closest cluster center

plt.plot(range(1,14),Array)

plt.title('Elbow Method')

plt.xlabel('Number of clusters')

plt.ylabel('Sum of Squares within groups')

plt.show()

# K-Means clustering algorithm on Rider age and Gender using Uber cabs

kmeans=KMeans(n\_clusters=4,init='k-means++',random\_state=0)

# fit function will give output of kmeans but fit\_predict will give the cluster index for each sample

Y=kmeans.fit\_predict(X)

plt.scatter(X[Y == 0,0], X[Y == 0,1],s=25,c='red',label='cluster 1') #s --> zoom level

plt.scatter(X[Y == 1,0], X[Y == 1,1],s=25,c='blue',label='cluster 2')

plt.scatter(X[Y == 2,0], X[Y == 2,1],s=25,c='magenta',label='cluster 3')

plt.scatter(X[Y == 3,0], X[Y == 3,1],s=25,c='cyan',label='cluster 4')

plt.scatter(kmeans.cluster\_centers\_[:,0],kmeans.cluster\_centers\_[:,1],s=25,c='yellow',label='Centroid')

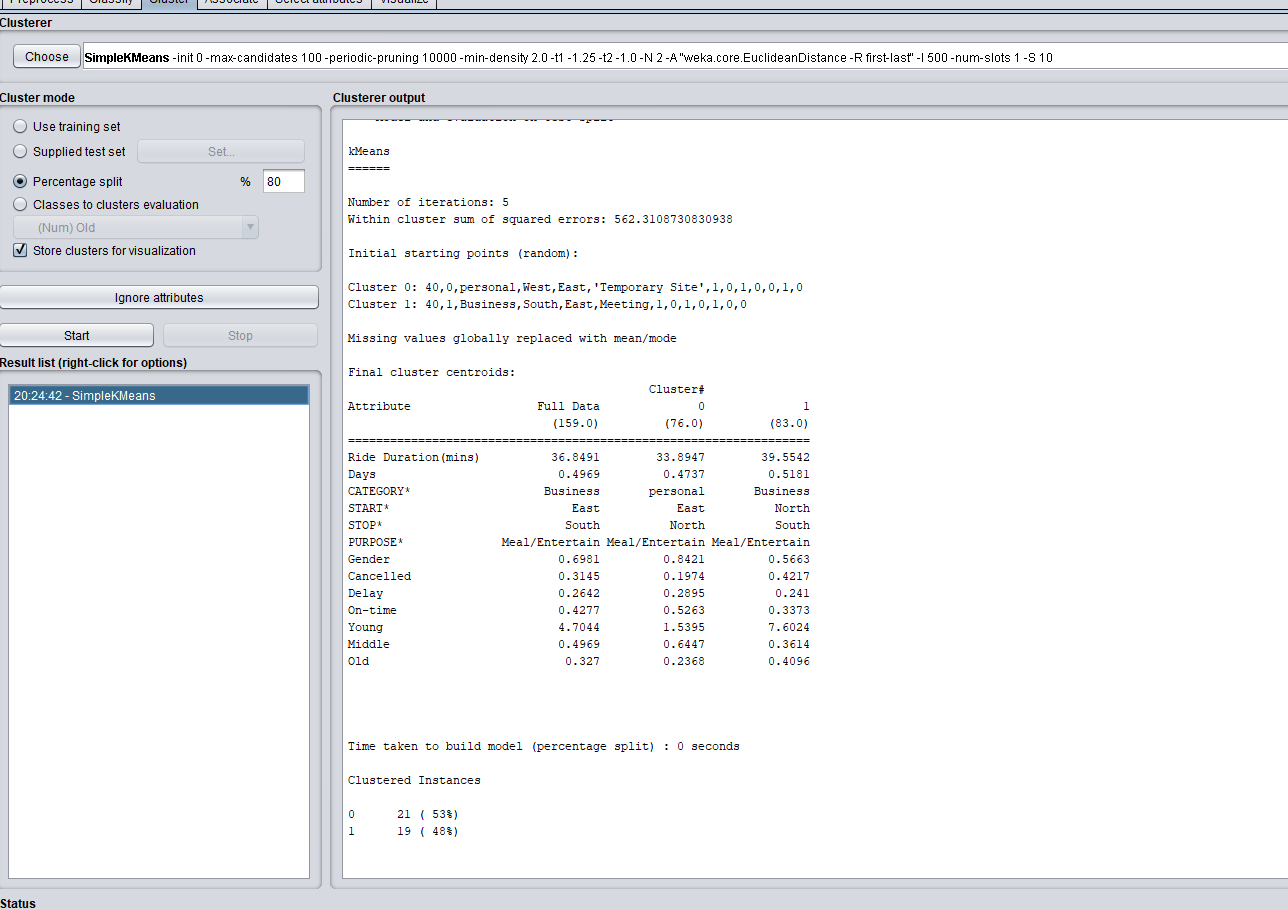
plt.title('K-Means Clustering')

plt.xlabel('Gender')

plt.ylabel('Rider Age')

plt.legend()

plt.show()

****

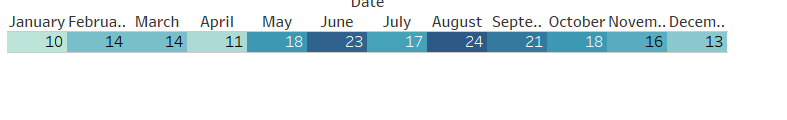
**9) To recommend the cabs based on drivers rating.**

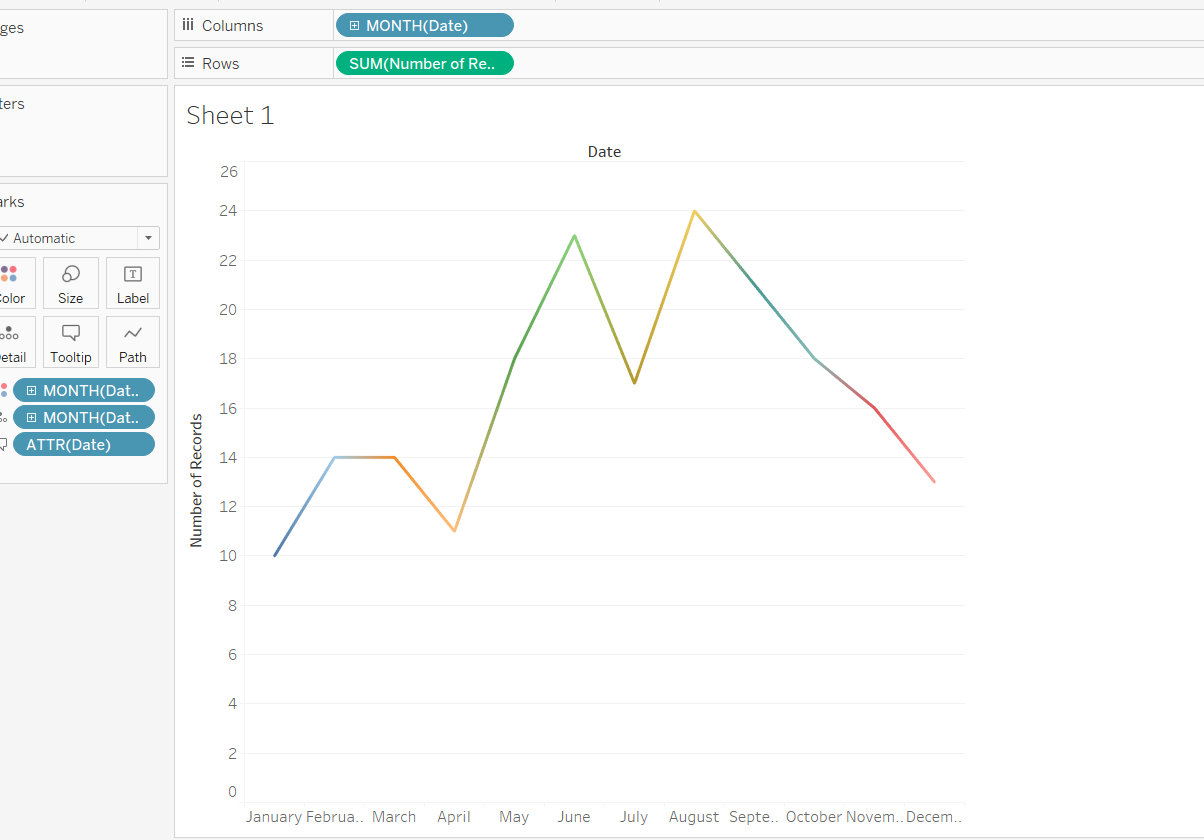
**?**

**Solution 9: It is recommender system problem which can be solved by using different traditional recommendation algorithms like Content Based Recommendation Algorithm, Collaborative Filtering Based Recommendation Algorithm and Hybrid Recommendation Algorithm.**

**10) To analyze most number of rides taken in a month.**

**Solution 10: To analyze the most number of rides in a month we can use analytical tools like tableau and Power BI. With the help of visualization tools we can provide different graphs for rides taken in a month.**

****

****