Ford Ka case study

Prof. N Mirbagheri

Mohammad Hadi Farahani

سوال :آیا بخش بندی بازار بر مبنای داده‌های روانشناختی (نیازها، و نگرشهای افراد) می‌تواند منتجر به شناسایی گروه‌های متمایزی از بازار شود؟ اگر این چنین باشد، این بخشهای بازار چه تمایزی از یکدیگر دارند و چگونه می‌توانند تفاوتهای Ka Choosers و Ka Non-Choosers را نشان دهند؟ (با تحلیل داده‌های ارائه شده و با استفاده از cluster analysis، در این رابطه توضیح دهید.)

برای پاسخ به این سوال نیاز داریم تا فرآیند کلاسترینگ را برای دیتاست جمع آوری شده از پاسخ دهندگان پیاده سازی نمائیم و بررسی کنیم که در تعداد بهینه ای از کلاستر، آیا تفاوت قابل توضیح و معناداری از لحاظ متغیرهای روانشناختی بین کلاسترها وجود دارد یا خیر.

ما به صورت کلی نیاز داریم تا برای تعیین کلاسترهای خود تعداد بهینه آن را در دیتاست خود پیدا کنیم. برای اینکار یک راه اینست که ابتدا از روش agglomerative استفاده نمائیم و بر اساس تغییرات فاصله درون کلاستری، استدلال مبتنی بر دندوگرام و نمودار elbow تعداد بهینه را تعیین کنیم و سپس از روش non-hierarchical یا k-means استفاده کنیم تا بر اساس تعداد بهینه کلاسترها از روش سلسله مراتبی، اعضای هر کدام را در بهینه ترین حالت تعیین کند و خروجی آن را به عنوان label در دیتاست وارد کنیم.

ابزارهای متفاوتی برای اجرای الگوریتم های کلاسترینگ وجود دارند که ما تلاش کردیم برای یادگیری خودمان از زبان پایتون و نرم افزار jupyter استفاده نمائیم. که در ادامه کدهارا توضیح مختصری داده و سپس درمورد نتایج خود بحث خواهیم کرد:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
```

ابتدا کتابخانه های لازم را وارد کردیم؛ که از numpy برای انجام محاسبات خطی، از pandas برای ساخت تیبل و مرتب سازی داده ها، از matplotlib.pyplot برای رسم نمودار های مختلف مثل اسکترپلات و sklearn برای اجرای الگورتیم های کلاسترینگ استفاده می گردد. سپس فایل دیتابیس خود را آپلود کردیم و دیتا را فراخوانی کردیم:

```
cust_df = pd.read_csv(r"C:Desktop\Book1.csv")
cust_df.head()
```

	Respondent Number	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	...	Q53	Q54	Q55	Q56	Q57	Q58	Q59	Q60	Q61	Q62
0	1	6	2	4	3	1	5	5	3	4	...	2	4	5	4	5	3	4	4	4	2
1	2	7	7	7	5	4	4	5	4	5	...	1	1	1	1	5	4	3	5	4	5
2	3	5	4	6	5	7	5	3	5	4	...	3	5	6	3	4	4	5	3	4	4
3	4	4	2	5	4	2	4	5	4	3	...	3	5	4	4	4	2	5	5	5	3
4	5	5	5	7	6	7	3	4	5	4	...	6	4	5	5	4	5	4	3	4	5

دیتاست ما قبل از انجام هرگونه محاسبات آماری از جمله کلاسترینگ نیاز به پردازیش اولیه دارد، لذا ابتدا ستون شماره پاسخ دهندگان را حذف کردیم، سپس داده های گمشده از دیتاست را پاک کردیم، همچنین جهت یکسان شده همه مقیاس ها داده هارا را استاندارد کردیم که میانگین صفر و انحراف معیار یک داشته باشند. البته لازم به ذکر است در این کیس خاص سوالات همه طیف ۷ تایی لیکرت هستند و نیازی به استاندارد سازی نبود ولی به عنوان یک فرآیند استاندارد و جلوگیری از اشتباهات احتمالی ما سعی میکنیم در تمام فرآیندهای تحلیل داده خودمان این موضوع را رعایت نمائیم:

```
df = cust_df.drop('Respondent Number', axis=1)
df.head()
```

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	...	Q53	Q54	Q55	Q56	Q57	Q58	Q59	Q60	Q61	Q62
0	6	2	4	3	1	5	5	3	4	4	...	2	4	5	4	5	3	4	4	4	2
1	7	7	7	5	4	4	5	4	5	5	...	1	1	1	1	5	4	3	5	4	5
2	5	4	6	5	7	5	3	5	4	5	...	3	5	6	3	4	4	5	3	4	4
3	4	2	5	4	2	4	5	4	3	4	...	3	5	4	4	4	2	5	5	5	3
4	5	5	7	6	7	3	4	5	4	2	...	6	4	5	5	4	5	4	3	4	5

```
# we change NaN to number for modeling
from sklearn.preprocessing import StandardScaler
X = df.values[:,:]
X = np.nan_to_num(X)

# this way, we standardize data for equalifiying
Clus_dataSet = StandardScaler().fit_transform(X)
Clus_dataSet
```

```
array([[ 0.58659625, -1.05009144, -0.34514518, ...,  0.16593271,
         0.24795029, -1.28828394],
       [ 1.23836986,  1.49867419,  1.98691686, ...,  0.89370777,
         0.24795029,  1.02323031],
       [-0.06517736, -0.03058519,  1.20956285, ..., -0.56184234,
         0.24795029,  0.25272556],
       ...,
       [ 0.58659625,  0.47916794, -1.1224992 , ..., -2.01739245,
        -1.30173903, -1.28828394],
       [-1.36872458, -0.54033831,  1.20956285, ...,  0.89370777,
         1.02279495, -0.51777919],
       [ 1.23836986,  0.98892106, -1.1224992 , ...,  0.89370777,
        -0.52689437,  1.79373506]])
```

یک نکته قابل ذکر است که ما به عنوان محقق ممکن است بتوانیم با استدلال های منطقی اقدام به حذف برخی از فیچرهای این دیتاست نمائیم ولی از آنجاییکه این فیچرها از سوالات استانداردی بدست آمده اند، حذف هر یک ممکن است باعث وزن دار شدن سایر فیچرها در مدل شود، به عبارتی حتی اگر بتوانیم اقدام به حذف متغیرها کنیم صلاحیت این کار را نداریم چون هر سوال فلسفه و دلیل اصلی وجودی ای دارد که ما لزوما آنرا نمی دانیم و لذا باید به همه وزن یکسانی در تحیل بدهیم. البته اگر در بحث انجام محاسبات اگر حذف برخی فیچرها باعث سریعتر شدن محاسبات می شد این امر قابل توجیح بود ولی چون از این نظر هم مشکلی نداشتیم سعی کردیم با یک نگاه unbiased به متغیرها داشته باشیم. یکی دیگر از تحلیل هایی که میتوانیم قبل از انجام کلاسترینگ داشته باشیم این است که متغیرهایی که احتمالا ناتوان از ایجاد تمایز در نمونه هستند را حذف کنیم. یکی از راهکارها اینست که انحراف از معیار متغیرهارا محاسبه کنیم و با این استدلال که مواردی انحراف معیار پایینی دارند احتمالا توان کمتری در ایجاد تمایز در نمونه دارند، برخی سوالات را حذف کنیم. البته اگرچه این رویکرد با دیدگاه unbiased بودن در تضاد است ولی ما انحراف معیار تمام سوالات را بررسی کردیم. لازم به ذکر است ما تفاوت معانداری در انحراف معیار ها دارشتیم. رنج انحراف معیار متغیرها قابل مشاهده است ولی به صورت کلی معیار خاصی وجود ندارد که ما تاچه انحراف معیاری را حذف کنیم و تا حدی پایینتر از انحراف معیاری را حذف کنیم. مثلا تنها دو متغیر با انحراف معیار زیر ۱ داشتیم (Q7, Q8) که مثلا این دو سوال به مسائلی مربوط به هندلینگ و قابلیت اتکای خودرو می پردازد که ممکن است به ماشین های کوچک مرتبط باشد و ما نمیتوانیم فقط با اتکای به انحراف معیار پایین آنها را حذف کنیم.

MAX STDEV	2.09806183
MIN STDEV	0.96609178
RANGE	1.13197005

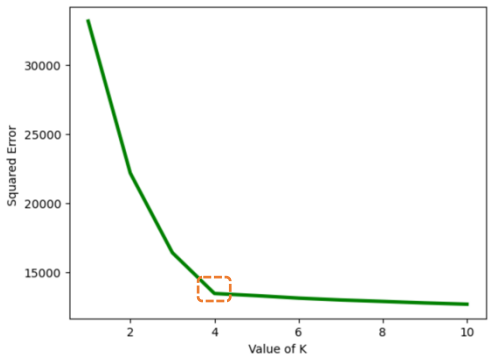
پس به صورت کلی ما تصمیم گرفته ایم بدون حذف هیچ متغیری جلو برویم تا با نگاه unbiased بتوانیم تحلیل خود را ارائه دهیم و در نهایت هنگام تحلیل تصمیم بگیریم که آیا میانگین متغیر بین کلاستر ها اینقدر متفاوت است که باید تحلیل و توصیف کاراکتر استفاده شود یا خیر.

حال برای تعیین تعداد بهینه کلاستر یک خلاقیت بخرج دادیم و مقدار کدهای استفاده شده را کاهش داده ایم. از الگوریتم k-means استفاده کردیم و گفتیم به ازای ۱ تا ۱۱ کلاستر مجذور فاصله درون گروهی را حساب کند و در حقیقت کاری که الگوریتم hierarchical به صورت نرمال انجام میداد را برای تعداد بهینه ای از کلاسترها (چون برای این تعداد دیتا بیشتر از ۱۱ تا کلاستر منطقی نبود) از روش k-means، که بهینه ترین کلاسترهارا استخراج میکند، انجام دادیم و SE را به ازای هر یک حساب کردیم. طبیعتا بیشتر از ۱۱ کلاستر ارزش بررسی نداشت و نیاز نبود مقدار SE برای بیشتر از آن محاسبه شود (مشابه کاری که الگوریتم سلسله مراتبی به صورت پیشفرض انجام میدهد):

```
# finding elbow (the best k for k-means)
SE =[]
for i in range(1, 11):

    KM = KMeans(n_clusters = i, max_iter = 500)
    KM.fit(X)

# calculates squared error for the clustered points
    SE.append(KM.inertia_)
```

حال خروجی مدل که فاصله درون کلاستری می شود را به ازای هر تعداد کلاستر تعیین شده در غالب نمودار elbow فراخوانی میکنیم.

```
# plot the cost against K values
plt.plot(range(1, 11), SE, color ='g', linewidth ='3')
plt.xlabel("Value of K")
plt.ylabel("Squared Error ")
plt.show() # clear the plot

# the point of the elbow is the
# most optimal value for choosing k
```



همانطور که مشاهده میشود شیب تغییر فاصله درون کلاستری[1] تا تعداد ۴ کلاستر زیاد بوده و پس از آن به شدت تغییر میکند (کاهش می یابد)، لذا میتوان نتیجه گرفت تعداد ۴ کلاستر از لحاظ آماری بهینه خواهد بود. حال در روش k-means تعداد ۴ کلاستر را میدهیم و میخواهیم الگوریتم با خزش سنترویید درون فضای ۶۲ بعدی داده ها آن هارا به بهینه ترین حالت به ۴ کلاستر مجزا تقسیم نماید. همانطور که مشخص است خروجی داده شده است که درآن به ترتیب لیبل هر رکورد را نشان داده است:

```
k_means = KMeans(init = "k-means++", n_clusters = 4, n_init = 200 , max_iter = 500)
k_means.fit(X)
labels = k_means.labels_
print(k_means.inertia_)
print(labels)
```

```
13440.095160256411
[1 0 2 1 2 0 0 0 3 1 0 2 3 1 0 1 0 1 0 3 2 2 0 2 2 3 2 1 1 1 2 0 1 1 2 2 2
 2 0 1 0 1 2 1 2 2 1 1 2 1 1 3 2 2 3 3 1 2 0 2 1 1 0 1 3 2 0 0 1 0 0 1 3 2
 0 1 1 2 1 1 0 2 0 0 2 0 1 0 3 2 1 3 2 1 2 0 1 1 0 3 2 2 2 0 3 0 1 2 0 0 2
 3 1 0 0 2 1 0 0 0 0 1 0 2 3 2 1 0 1 2 2 3 0 3 1 0 2 1 2 1 1 0 1 1 3 2 3 0
 0 0 0 1 0 0 0 3 0 0 1 1 1 2 0 0 0 0 0 2 0 3 0 3 0 0 2 0 1 1 2 2 3 1 2 3 2
 0 2 2 1 1 2 2 0 0 2 2 3 1 1 1 2 1 2 3 1 2 1 1 3 0 1 0 3 1 1 2 0 1 1 0 2 0
 2 0 0 0 0 3 1 1 1 1 3 1 2 1 1 0 0 3 1 0 2 1 0 0 2 3 2 0]
```

حال لیبل خروجی از الگوریتم را در یک ستون جدید با عنوان clus_km وارد دیتاست میکنیم:

```
df["Clus_km"] = labels
df.head(5)
```

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	...	Q54	Q55	Q56	Q57	Q58	Q59	Q60	Q61	Q62	Clus_km
0	6	2	4	3	1	5	5	3	4	4	...	4	5	4	5	3	4	4	4	2	1
1	7	7	7	5	4	4	5	4	5	5	...	1	1	1	5	4	3	5	4	5	0
2	5	4	6	5	7	5	3	5	4	5	...	5	6	3	4	4	5	3	4	4	3
3	4	2	5	4	2	4	5	4	3	4	...	5	4	4	4	2	5	5	5	3	1
4	5	5	7	6	7	3	4	5	4	2	...	4	5	5	4	5	4	3	4	5	3

سوال ما اینست که آیا این کلاسترها از لحاظ متغیرهای روانشناختی متمایز می باشند و ایا قابل تفسیر هستند یا خیر؟ یک روش اینست که ما میانگین تمام متغیرهای روانشناختی را برای هر کلاستر بدست بیاوریم. با مقایسه میانگین هر متغیر برای هر کلاستر میتوانیم متوجه شویم آیا تفاوت معنی داری دارند و آیا مقادیر آنها بینش خاصی نسبت به تفاوتشان یا پرسونا شان به ما میدهد؟

[1] در اینجا مفهوم SE همان مجذور فاصله دورن کلاستری می باشد.

```
df.groupby('Clus_km').mean()
```

Clus_km	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	...	Q53	Q54	Q55	Q56
0	3.973333	1.973333	4.013333	4.080000	1.760000	3.933333	3.933333	3.946667	3.920000	3.933333	...	4.240000	4.080000	4.146667	3.866667
1	6.512821	6.512821	3.884615	4.025641	4.012821	3.974359	4.025641	3.756410	4.076923	4.051282	...	1.538462	1.461538	1.346154	1.384615
2	4.015385	3.769231	5.938462	6.015385	6.015385	3.969231	3.800000	4.107692	3.769231	3.815385	...	3.984615	3.800000	3.861538	3.953846
3	6.500000	3.562500	3.781250	1.500000	3.937500	4.218750	3.562500	3.843750	3.718750	3.750000	...	6.468750	3.843750	3.656250	4.125000

قبل از شروع توصیف کلاسترها بهتر است بررسی کنیم هر کلاستر چند درصد از کل تعداد جامعه را به خود اختصاص داده است:

Row Labels	Count of Count
0	31.20%
1	30.00%
2	26.00%
3	12.80%
(blank)	0.00%
Grand Total	**100.00%**

همانطور که مشخص است کلاستر ۰ و ۱ با یک سهم تقریبا یکسان هر کدام ۳۰ درصد جامعه را به خود اختصاص داده اند و کلاستر ۳ کوچکترین کلاستر موجود می باشد.

حال به توصیف کلاسترهای بدست آمده می پردازیم:

1. کلاستر شماره ۰ (easy going and functionality): دغدغه تعمیرات را نمی خواهند داشته باشند (Q3). دنبال یک راهکار ساده و ارزان برای حمل و نقل هستند (Q4, Q33) و پرستیژ خودرو برایش اهمیت ندارد. عملکرد فنی فوق العاده خودرو مثل سرعت شتاب بالا برایش اولویت ندارد (Q33). صرفا خودرویی میخواهد که راحت باشد و کارش را راه بندازد (Q31, Q34) و صرفا سفر درون شهری دارند (Q32). کانسپت خودروهای کوچک را درک نکرده اند و خیلی برایشان ارزشی قائل نیستند و از منظر اجتماعی استفاده از خودروهای بزرگ را ارزش میدانند و از وضعیت فعلی بازار خودرو و ارضا نیازهایش رضایت دارد (Q39-42). کاراکتر و خاص بودن خودرو برایشان مهم نیست (Q44). دوست ندارد با middleman ها معامله کند و احتمالا اینطور خودرو را ارزانتر می خرد (Q43).

   به صورت کلی کلاستر ۰ به دنبال عملکرد و راحتی در استفاده هستند و دنبال دردسر نمی گردند و ضرورتی برای استفاده از خودروهای کوچک نمی بنند، این گروه ۳۱,۲ درصد از کل جامعه را تشکیل میدهند. طبق شیت Cluster-Preference ۴۳,۶ درصد این کلاستر از دسته ka chooser ها و ۳۹,۷ درصد از دسته ka non-chooser ها بوده اند.

2. کلاستر شماره ۱ (big car as a home): دوست دارند خودرویشان ظاهر مدرن و بروزی داشته باشد (Q1, Q57). خودرو بیش از یک وسیله حمل و نقل ساده برایشان است و با آن ارتباط خیلی نزدیکی دارند (Q4, Q52) و قابلیت اعتماد و کیفیت بالا برایشان مهم است و برای نگهداری خودرویش وقت می گذارد (Q53, Q59). سفرهای طولانی و بلند دارند و دوست دارند خودرویشان حداکثر تجهیزات ممکن را داشته باشد و در سفر راحت باشد (Q14, Q15, Q31). درمورد تاثیر خودرو روی محیط زیست اهمیتی قائل نیست (Q16). عملکرد فنی و سرعت و شتاب خودرو برایش مهم است (Q17, Q37). وفاداری به برند بالایی ندارد (Q18). ایرودینامیک خودرو برایشان مهم است (Q21). خودروهای کوچک را به اندازه کافی ایمن می دانند (Q22). خودروهای کوچک را به اندازه کافی ایمن می دانند و می گویند لزوما دیگر برای افراد کوچک نیست (Q22, Q26)، از لحاظ اجتماعی خودروهای بزرگ را مناسب فضای امروزی میدانند (Q41) از وضعیت فعلی بازار خودرو و ارضا نیازهایش رضایت دارد (Q42). خیلی خودشان را در مسائل مربوط به خودرو متخصص نمی دانند (Q25). وطن پرست هستند و تولید داخل را دوست دارند (Q28) ولی در خریدهایش محصولات را فارغ از مبدا تولید بررسی میکند. خودرو را یکی از ابزارهای ابراز آزادی و استقلال میداند (Q58). خیلی موافق حمل و نقل عمومی نیست و با تبعیض قیمتی بین خودروهای کوچک و بزرگ موافق نیست و دادن مالیات بیشتر روی خودروهای بزرگ را نمی پسندد (Q62).

به صورت کلی کلاستر ۱ ارتباط عاطفی عمیقی بین صاحب خودرو با وسیله اش وجود دارد، به عبارتی خودرو را به مثابه خانه خود می دانند و اهل خودروهای بزرگ هستند، این گروه ۳۰ درصد از کل جامعه را تشکیل میدهند. طبق شیت Cluster-Preference ۴۶,۷ درصد این کلاستر از دسته ka chooser ها و ۲۲,۶۷ درصد از دسته ka non-chooser ها بوده اند.

3. کلاستر شماره ۲ (Patriot trendies): ترند بودن خودرو و ظاهر بروز برایشان مهم است و دوست دارند همیشه آخرین طراحی و مدل خودرو و طراحی داخلب منحصر بفردی را داشته باشند (Q1, Q2, Q46, Q48). سفر های طولانی نمی روند (Q14). برای خودرو بیشتر از بودجه مالیشان خرج نمیکنند و اقتصادی فکر میکنند (Q20). آپشن های فایننسینگ و تسهیلات مالی تاثیری رو تصمیم خریدشان ندارد و احتمالا نباید نقطه تمرکز ما برای این سگمنت باشد (Q23). از خودرو خود فقط در حمل و نقل شهری استفاده میکنند (Q24). راحتی خودرو برایشان مهم نیست (Q31)! از منظر اجتماعی استفاده از خودروهای بزرگ را مناسب نمی بیند و خودرو را وسیله ای برای ابزار خصوصیات خودش میبیند (Q41, Q44, Q45). در هنگام خرید خودرو احساسات غلبه زیادی بر منطق آنها دارد، خودروهای سبک مورد علاقه خود را غالبا از برندهای خارجی و غیرملی میداند و معتقد است باید با خرید خودروهای داخلی به پیشرفت آنها کمک کرد (Q47, Q49, Q50). تجهیزات کامل در خودرو اولیت زیادی برایشان ندارد و ظاهر بیشتر برایشان مهم است (Q51). وفاداری خاصی به ماشینش ندارد و میتواند سریع سوییچ کند (Q52). کیفیت و پایداری خودرو اولویت آنها نیست (Q53). معتقد است امروز ظاهر خودروها خیلی متنوع شده است، معتقد است سبک جدید خودروهای کوچک روی عملکردشان تاثیر قابل توجهی نداشته است و به اندازه کافی خوب هستند (Q55, Q56).

به صورت کلی کلاستر ۲ طراحی و ظاهر ماشین برایشان مهم است ولی امکانات و کیفیت آن یا ضرباتی که محیط زیست میزند برایشان مهم نیست و وفاداری به برند خاصی نیستند، این گروه ۲۶ درصد از کل جامعه را تشکیل میدهند. طبق شیت Cluster-Preference ۴۴,۶ درصد این کلاستر از دسته ka chooser ها و ۴۹,۲۳ درصد از دسته ka non-chooser ها بوده اند. به عبارتی ford ka برای این گروه یا بسیار مطلوب بوده یا از آن خوششان نمی آمده است. که مشخص است چون فورد تلاش کرده بود امکانات زیادی را در ماشین قرار دهد و در عین حال ظاهری بسیار هنجار شکنانه و مدرن به آن بدهد که با خواسته های آنان تناقض دارد.

4. کلاستر شماره ۳ (act loyally ,care environmentally): خیلی به مد توجهی ندارند (Q2). خودروهای کوچک را باپرستیژ میدانند(Q5). از خودروهایشان برای سفرهای طولانی استفاده میکنند (Q14). دوست دارند خودرویشان تجهیزات کاملی داشته باشند (Q15). محیط زیست و تاثیری که خودرویش روی آن میگذارد برایش اهمیت دارد (Q16, Q19). چالاک و سریع بودن خودرو برایشان مهم است، وفاداری به برند بالایی دارند و بودجه بندی را در تهیه خودرو به دقت رعایت میکنند، ایرودینامیک خودرو برایش مهم است و امنیت خودرو های کوچک را مناسب میداند (Q17, Q18, Q20-22). روش های فایننسینگ برایشان جذابیت دارد و میتواند روی خریدشان تاثیر بگذارد، خود را در حوزه خودرو صاحب نظر نمیداند، خودروهای کوچک را مناسب عام میداند، تعصبی به تولید داخل ندارد و خودرو را یک وسیله کاربردی برای روزمره میداند (.(Q23-30

به صورت کلی کلاستر ۳ هم بحث اقتصادی و بودجه بندی برایشان مهم است و هم به محیط زیست اهمیت زیادی می دهند، این گروه ۱۲,۸ درصد از کل جامعه را تشکیل میدهند. طبق شیت Cluster-Preference ۵۶ درصد این کلاستر از دسته ka chooser ها و ۱۲,۵ درصد از دسته ka non-chooser ها بوده اند.

Count of Count	Column Labels			
Row Labels	1	2	3	Grand Total
0	29.31%	18.06%	50.00%	31.20%
1	30.17%	31.94%	27.42%	30.00%
2	25.00%	44.44%	6.45%	26.00%
3	15.52%	5.56%	16.13%	12.80%
Grand Total	**100.00%**	**100.00%**	**100.00%**	**100.00%**

در pivot table فوق سطر شماره کلاستر و ستون وضعیت preference می باشد. می توان مشاهده کرد که ۵۰ درصد از کل ka non-chooser ها از گروه ۱ بوده اند. ۴۴,۴ درصد از middle-chooser ها که ما بهشان so-so میگوییم از کلاستر ۲ بوده اند و مجموعا ۶۰ درصد کل ka chooser ها از کلاستر ۰ و ۱ بوده اند.

برای اینکه بدانیم دقیقا چه ویژگی هایی بین گروه ka-chooser و ka non-chooser ها تمایز ایجاد میکند از یکone-way ANOVA تست استفاده میکنیم. در این روش مقایسه میکنیم که آیا میانگین دسته های مختلف تفاوت معناداری با هم دارند یا نه؟ این روش را برای متغیرهای demographic و psychographic انجام میدهیم و سپس بر اساس درجه significance تفاوت ها تصمیم میگیریم. البته استفاده از ANOVA فرض هایی دارد که مثلا توزیع داده ها نرمال باشد که البته بجز سن بقیه متغیرها کتگوریکال هستند بدون نرمال سازی ادامه میدهیم. البته راهکار ساده تر این بود که صرفا میانگین هارا محاسبه میکردیم و طبق آنها استدلال میکردیم ولی چون مبنای آماری ندارد بهتر است از این روش استفاده کنیم.

1 - متغیرهای روانشناختی: از مسیر Analyze> Compare means> One-way ANOVA می رویم و در لیست dependent متغیرهای روانشناختی را وارد میکنیم و در factor متغیر preference را وارد میکنیم. خروجی جدول را در اکسل قرار داده ایم و چون طولانی بود اینجا قرار ندادیم (Q2 clusters results.xlsx> ANOVA-PSY). با یک conditional formatting مشخص میکنیم کدام مقادیر در سطح ۰,۰۵ statistically sig هستند. همانطور که مشاهده می شود سوال های ۱ تا ۴ و ۱۲ و ۱۷ و ۲۰ و ۲۳ و ۲۴ و ۳۱ تا ۳۶ و ۳۸ تا ۴۱ و ۴۳ تا ۴۶ و ۴۸ تا ۵۳ و ۵۴ تا ۵۶ از لحاظ آماری تفاوت میانگین بین گروه های preference معنی دار است. محتوای این سوالات به مسائلی همچون ترند و فشن بودن، مصرف سوخت بهینه، چالاک وتند بودن سرعت خوردو، توجه به بودجه بندی و امکان فایننس برای خوردو، نحوه نگرش اجتماعی و شخصی نسبت به خودروهای کوچک، کیفیت و پایداری خودرو، میزان تجهیز بودن به امکانات و فیچرها و طراحی داخلی اشاره دارد. همان چیزی که پیشتر بحث کردیم یعنی ka chooser ها و ka non-chooser ها در مواردی که گفته شده تفاوت های قابل توجهی با یکدیگر دارند و همین موارد می تواند معیار تصمیم گیریشان باشد.

2 - متغیرهای دموگرافیک: از مسیر Analyze> Compare means> One-way ANOVA می رویم و در لیست dependent متغیرهای دموگرافیک را وارد میکنیم و در factor متغیر preference را وارد میکنیم. خروجی جدول را در اکسل قرار داده ایم و چون طولانی بود اینجا قرار ندادیم (Q2 clusters results.xlsx> ANOVA-DEMO). با یک conditional formatting مشخص میکنیم کدام مقادیر در سطح ۰,۰۵ statistically sig هستند. همانطور که مشاهده می شود بجز متغیرهای سن و 1st time purchase میانگین بین گروه ها در سایر متغیرها از لحاظ آماری معنادار نیست. البته میتوان گفت سن نیز یک پروکسی از متغیرهای روانشناختی می باشد. مثلا افراد با سن بالاتر دیدگاه های سخت گیرانه تری درمورد نگرش نسبت به خودروهای کوچک و کیفیت خودرو وپایداری عملکرد دارند ولی جوانان نگاه لیبرال تری نسبت به فشن و سرعت خودرو و بودجه بندی و امکان فایننسینگ دارند. لذا می توان ادعا کرد متغیرهای دموگرافیک بهترین معیاری که میتوان برای دسته بندی گروه های ka chooser و ka non-chooser ها استفاده کرد نمی باشد و توان ایجاد تمایز بین آنها را به کیفیتی که متغیرهای روانشناختی انجام می دهند، ندارد.

جمع بندی آنکه به نظر می رسد از لحاظ آماری متغیرهای روانشناختی توان خوبی در ایجاد تمایز بین گروه های ka chooser و ka non-chooser داشته باشد و متغیرهای دموگرافیک از این امر ناتوان هستند. از لحاظ مفهومی و استدلالی نیز بحث کردیم هر گروه تشکیل شده از طریق متغیرهای روان شناختی توانسته اند معناداری و کاراکتر خاص خود را داشته باشند. همچنین به تفاوت های اصلی بین دو گروه از لحاظ متغیرهای روان شناختی در بخش ANOVA بحث کردیم و در بخش توضیح کلاسترها گفتیم که هر گروه دقیقا چه ویژگی های روانشناختی شاخصی دارد که میتواند در بحث targeting بسیار کارآمد باشد ولی چون بحث سوال ما نمی باشد به جهت کوتاه نگه داشتن متن به آن نپرداختیم.

- استفاده از روش dimension reduction برای کلاسترینگ (محاسبات اختیاری):

در ابتدای سوال ۲ بحث کردیم که ما برای کلاسترینگ باید از ۶۲ متغیر استفاده نمائیم. این تعداد متغیر بسیار زیاد بوده و در چنین مواقعی می توان از dimension reduction استفاده نمود تا تعداد فیچرهای دخیل در فرآیند کلاسترینگ را کاهش داد و در نتیجه دقت مدل را افزایش داد.

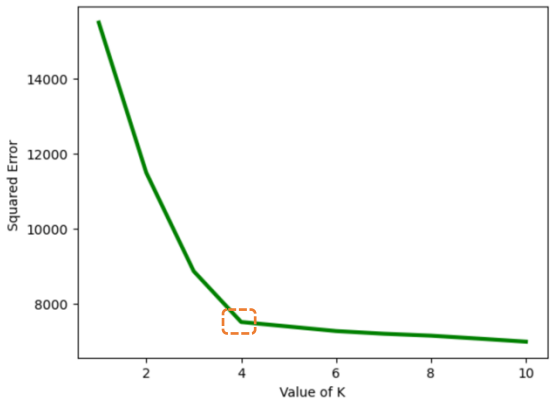
ابتدا بر اساس فایل (dimentionreduction component.ipynb) دیتاست اصلی را وارد کردیم و پس از تمیز سازی داده و استاندارد سازی، از عملگر PCA استفاده کردیم که به ازای تعداد کامپوننت های مختلف مقدار واریانس توضیح داده را نمایش دهد. به عبارتی هرچه تعداد کامپوننت های بیشتری داشته باشیم واریانس بیشتری از رفتار متغیرها توضیح داده می شود (اگر تعداد کامپوننت و فیچر یکی شود عملا ۱۰۰ درصد واریانس داده هارا توضیح میدهیم) ولی ابعاد مسئله نیز بیشتر می شود و یک تریدآف بین این دو وجود دارد و حد بهینه طبق نظر خبرگان در مسائل MLاینست که ۶۰ درصد variance explained کفایت میکند و ما با ۷ کامپوننت به این مقدار دست خواهیم یافت. لذا بجای ۶۲ فیچر که ۱۰۰ درصد واریانس را توضیح میدهند تنها از ۷ کامپوننت Dem Reduc استفاده خواهیم کرد و ۶۱,۷۸ درصد از واریانس داده هارا می توانیم توضیح دهیم که حد خوبی است:

```
# Calculate the variance

for i in range(2 , 20):
    pca = PCA(n_components= i)
    pca.fit(Clus_dataSet)
    print(f"{round(sum(list(pca.explained_variance_ratio_))*100, 2)} Total Variance Explained for {i} components ")
```

```
44.17 Total Variance Explained for 2 components
53.51 Total Variance Explained for 3 components
55.88 Total Variance Explained for 4 components
57.94 Total Variance Explained for 5 components
59.9 Total Variance Explained for 6 components
61.78 Total Variance Explained for 7 components
63.6 Total Variance Explained for 8 components
65.23 Total Variance Explained for 9 components
66.81 Total Variance Explained for 10 components
68.26 Total Variance Explained for 11 components
69.63 Total Variance Explained for 12 components
70.95 Total Variance Explained for 13 components
72.23 Total Variance Explained for 14 components
73.46 Total Variance Explained for 15 components
74.65 Total Variance Explained for 16 components
75.81 Total Variance Explained for 17 components
76.94 Total Variance Explained for 18 components
77.98 Total Variance Explained for 19 components
```

حال که متوجه شدیم ۷ کامپوننت عدد بهینه ای است، مقادیر آنهارا فراخوانی کرده و به عنوان دیتاست ورودی وارد مراحل کلاسترینگ میکنیم (مقادیر بدست آمده در dimention reduction- component.csv موجود است). ادامه مراحل کلاسترینگ در فایل (dimentionreduction component.ipynb) موجود است. تمامی مراحل گفته شده در بخش اول سوال را تکرار میکنیم و نمودار elbow را رسم میکنیم:

```
# plot the cost against K values
plt.plot(range(1, 11), SE, color ='g', linewidth ='3')
plt.xlabel("Value of K")
plt.ylabel("Squared Error ")
plt.show() # clear the plot

# the point of the elbow is the
# most optimal value for choosing k
```

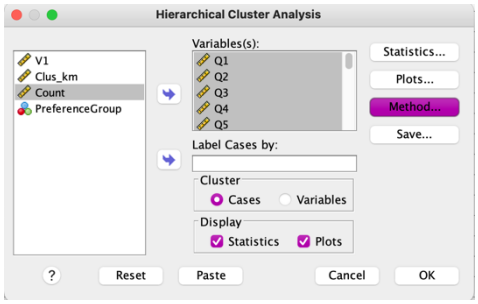


همانطور که مشخص است تعداد کلاستر بهنیه در این روش نیز ۴ عدد می باشد (نکته آنکه مقدار SE یا مجذور فاصله درون کلاستری در این روش نصف شده است).

حال خروجی این روش یعنی لیبل کلاستر هر رکورد را در یک فایل CSV ذخیره میکنیم. همچنین میانگین مقادیر هر کلاستر برای سوال ۶۲ گانه را مجدد محاسبه میکنیم. میخواهیم بررسی کنیم آیا این روش کلاسترهای متفاوت با تفسیر پذیری متفاوتی از روش قبلی ایجاد کرده است یا هر دو به نتایج یکسانی دست می یابند؟

همانطور که در فایل (Q2 clusters results.xlsx) مشخص است ما لیبل هر دو روش ۶۲ فیچر و دیمنشن ریداکشن را در ستون های BL و BM قرار داده ایم. در نگاه اول لیبل کلاسترها تغییر کرده است و برخی اعضا به کلاسترهای متفاوتی از روش ۶۲ فیچر تخصیص داده شده اند ولی با محاسبه میانگین ها متوجه میشویم نتایج یکسان است و فقط اعضای کلاستر ۱ و ۲ و ۳ به ترتیب در روش دیمنشن ریداکشن به اعضای کلاسترهای ۲ و ۳ و ۱ تبدیل شده اند ولی میانگین های آنها ثابت مانده است لذا می توان گفت هر دو روش به نتیجه یکسانی در کلاسترینگ دست یافته اند و نتایج قبلی ما robust هستند.

- استفاده از روش hierarchical برای تایید نتایج ۴ کلاستر:



اگرچه دو روش قبلی هر دو به ۴ کلاستر دست یافتند ولی جهت تمرین خودمان از روش سلسله مراتبی نیز تست کردیم که آیا نتایج قبلی مارا تائید میکند یا خیر؟ از منوی Analyze> classify> hierarchical clustering پلات های دندگرام و ایسیکل را فعال کردیم و از متد لینکیج استفاده کردیم. بدلیل بزرگ بودن نتایج امکان ارائه انها در فایل ورد وجود ندارد ولی خروجی نمودارهارا در فایل (Output1-dendogram.spv) موجود است که با ایجاد یک خط فرضی در دندوگرام و نمودار ایسیکل میتوانیم متوجه شویم تا جایی که ۴ کلاستر دست می یابیم فاصله درون کلاستری خیلی زیاد است ولی اگر بخواهیم به بیشتر از ۴ کلاستر مثلا ۵ تا دست یابیم فاصله ها خیلی کمتر می شود و لذا این موضوع نیز نتایج قبلی را تایید میکند.