گزارش پروژه داده کاوی

مریم واقعی

اطلاعات گزارش	چکیده
تاریخ: ۱۲ تیر ۱۴۰۲	تحلیل مشتریان یک شرکت با استفاده از داده کاوی یکی از مسائل مهم در بازاریابی است. در این پروژه، با استفاده از داده های موجود در فایل marketing_champain.csv، به تحلیل و بررسی مشتریان یک شرکت پرداخته شده است. با استفاده از این داده ها، شرکت می تواند نیازها، رفتارها و انواع مختلف مشتریان را بهتر درک کرده و محصولات خود را مدیریت کند. در این پروژه از روش های داده کاوی و مدل سازی استفاده شده است تا بهترین راهکارهایی برای بازاریابی و فروش محصولات به مشتریان پیشنهاد شود.
واژگان کلیدی: داده کاوی پیش پردازش تحلیل EDA الگوهای پرتکرار قوانین انجمنی ابرکلمات	

فهرست مطالب

۱-مقدمه ۳

۲- فاز اول پروژه ۳

۲-۱- شناخت و پیش پردازش داده ها ۳

۲-۱-۱- بخش اول ۳

۲-۱-۲- بخش دوم ۳

۲-۱-۳- بخش سوم ۱۲

۲-۱-۴- بخش چهارم ۱۳

۲-۱-۵- بخش پنجم ۱۵

۲-۲- تحلیل EDA ۱۵

۲-۲-۱- بخش اول ۱۵

۲-۲-۲- بخش دوم ۱۶

۲-۲-۳- بخش سوم ۱۶

۲-۲-۴- بخش چهارم ۱۷

۲-۲-۵- بخش پنجم ۱۸

۲-۲-۶- بخش ششم ۱۹

۲-۳- تسک های نمره اضافه ۲۳

۲-۳-۱- بخش اول ۲۳

۲-۳-۲- بخش دوم ۲۴

۳- فاز دوم پروژه ۲۷

۳-۱- تحلیل الگو و ارائه مدل های پیش بینی ۲۷

۳-۱-۱- بخش اول ۲۷

۳-۱-۲- بخش دوم ۲۸

۳-۱-۳- بخش سوم ۳۱

۳-۱-۴- بخش چهارم ۳۲

۳-۱-۵- بخش پنجم ۳۴

۳-۲- تسک های نمره اضافه ۳۵

۳-۲-۱- بخش اول ۳۵

۳-۲-۲- بخش دوم ۳۸

۱-مقدمه

امروزه در بازاریابی، شناخت و درک مشتریان از مسائل حیاتی برای شرکت ها به شمار می روند. در این راستا، تحلیل داده های مشتریان با استفاده از روش های داده کاوی یکی از مهمترین ابزارها برای بهبود فرآیند بازاریابی و فروش محصولات است. در این پروژه، با استفاده از داده های موجود در فایل marketing_champain.csv، به تحلیل و بررسی مشتریان یک شرکت پرداخته شده است. در این پروژه، ابتدا با استفاده از روش های پیش پردازش داده ها، داده های بدون ارزش و نویز را حذف و داده های مورد نیاز برای تحلیل را استخراج کرده ایم. سپس با استفاده از روش های داده کاوی و مدل سازی، الگوهای رفتاری و نیازهای مشتریان شناسایی و انواع مختلف مشتریان مدیریت شده اند. در نهایت، با توجه به نتایج به دست آمده، راهکارهایی برای بهبود بازاریابی و فروش محصولات به مشتریان پیشنهاد میشود.

۲- فاز اول پروژه

۲-۱- شناخت و پیش پردازش داده ها

۲-۱-۱- بخش اول

برای این بخش، متدی به نام attributes_information(df) نوشته شده است که یک dataframe را دریافت کرده و برای ستون های عددی آن مقادیر زیر را محاسبه میکند:

۱. Range: در صورتی که min < Q1-1.5IQR در این صورت بین محدوده [min, Q1-1.5IQR] داده پرت داریم فلذا کران پایین بازه مقادیر برابر با Q1-1.5IQR می باشد. اگر min > Q1-1.5IQR در این صورت در ابتدای بازه همه اعداد در محدوده مناسب هستند در نتیجه کران پایین بازه مقادیر برابر با min می شود.

برای کران بالا نیز اگر max > Q3 + 1.5IQR در این صورت در محدوده [max, Q3 + 1.5IQR] داده پرت داریم فلذا کران پایین بازه مقادیر برابر با Q3 + 1.5IQR است و در غیر این صورت داده ها در محدوده مناسب هستند فلذا کران بالای بازه مقادیر برابر با max میشود.

۲. Min: با استفاده از متد ()min از کتابخانه pandas محاسبه میشود.

۳. Max: مشابه min با استفاده از متد ()max از کتابخانه pandas محاسبه میشود.

۴. Mean: با توابع آماده pandas

۵. Median: با توابع آماده pandas

۶. Mode: با توابع آماده pandas

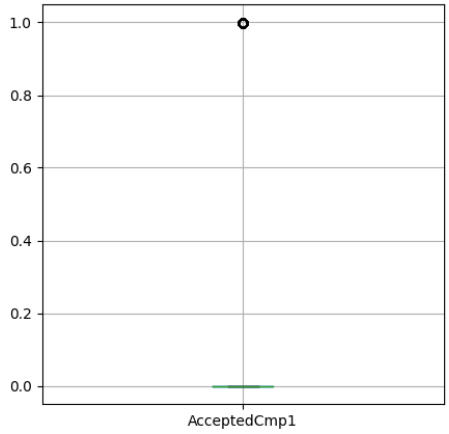
۷. Outlier_Range: مشابه ستون Range در صورتی که بازه ای از مقادیر پرت شناسایی شوند آن بازه را در این بخش قرار میدهیم که میتواند شامل یک یا هر دو بازه ی [min, Q1-1.5IQR] و [max, Q3 + 1.5IQR] باشد.

۸. Std: انحراف معیار نیز با استفاده از توابع آماده pandas قابل محاسبه است.

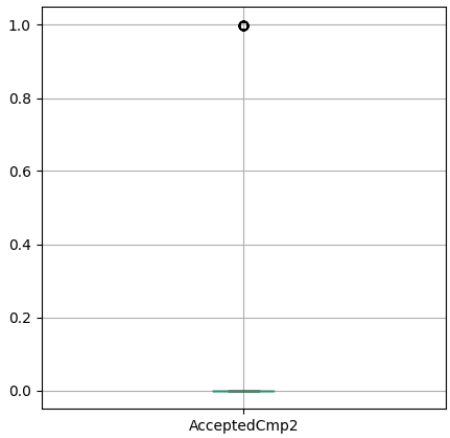
وقتی همه این مقادیر را برای ستون های عددی محاسبه کردیم و آنها را در یک dataframe قرار دادیم، این اطلاعات را در فایل attributes_information.csv ذخیره میکنیم که در کنار فایل گزارش قرار داده شده و میتوانید آن را مشاهده کنید.

۲-۱-۲- بخش دوم

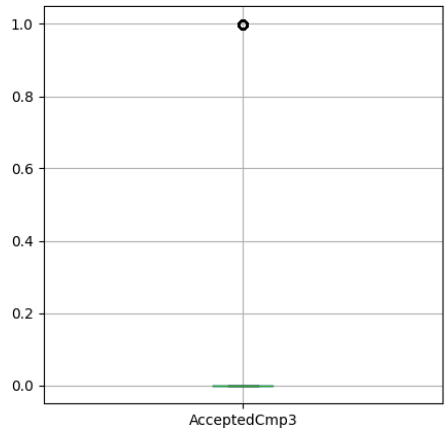
نمودار box plot مربوط به ستون های عددی با استفاده از متد boxplot در کتابخانه pandas رسم شده است که به صورت زیر میباشد:



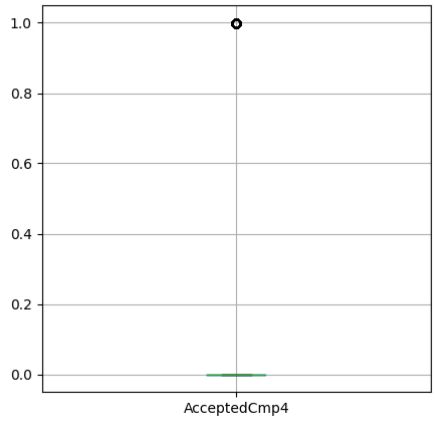
نمودار ۱- نمودار box plot مربوط به ستون AcceptedCmp1



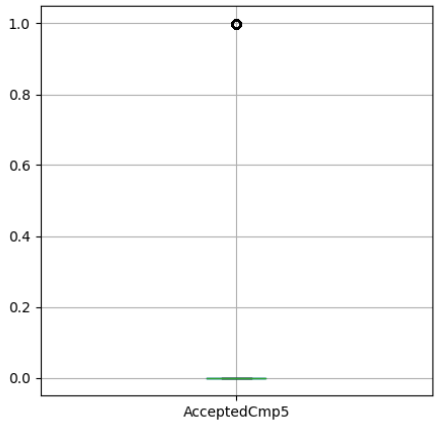
نمودار ۲- نمودار box plot مربوط به ستون AcceptedCmp2



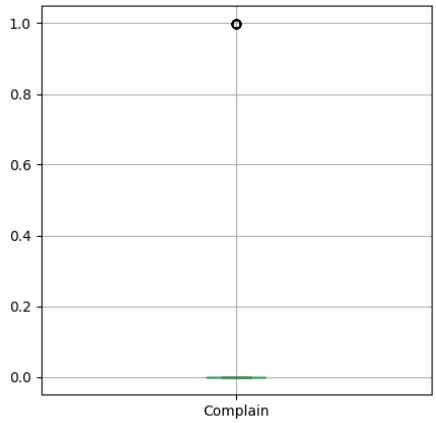
نمودار ۳- نمودار box plot مربوط به ستون AcceptedCmp3



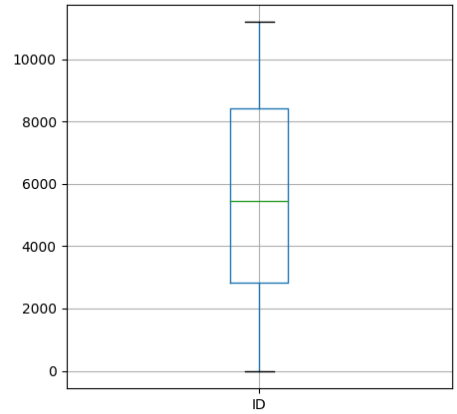
نمودار ۴- نمودار box plot مربوط به ستون AcceptedCmp4



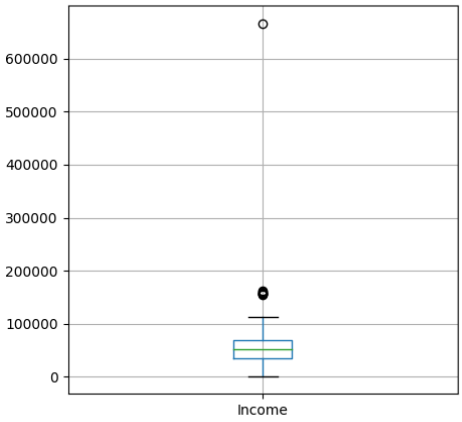
نمودار ۵- نمودار box plot مربوط به ستون AcceptedCmp5



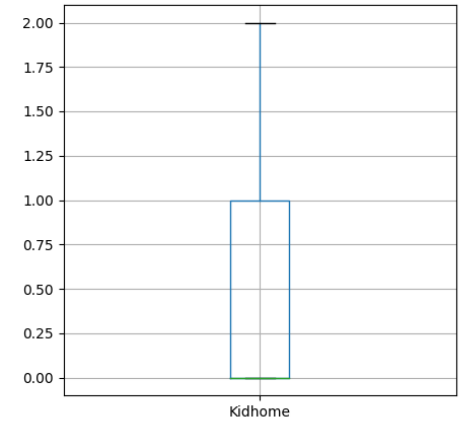
نمودار ۶- نمودار box plot مربوط به ستون Complain



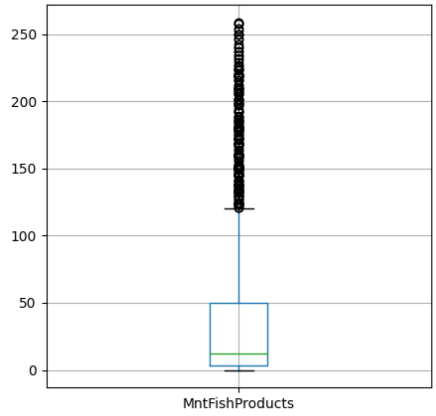
نمودار ۷- نمودار box plot مربوط به ستون ID



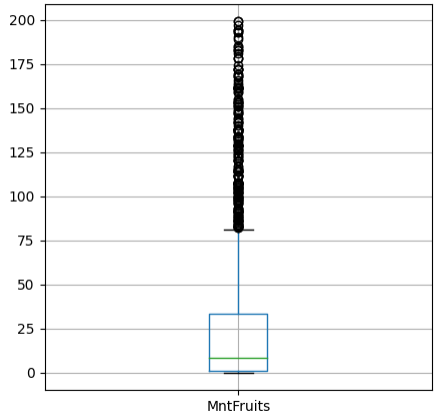
نمودار ۸- نمودار box plot مربوط به ستون Income



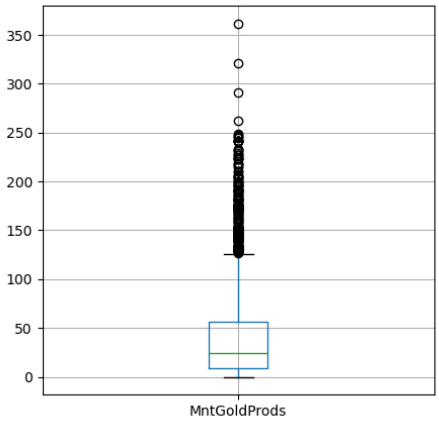
نمودار ۹- نمودار box plot مربوط به ستون Kidhome



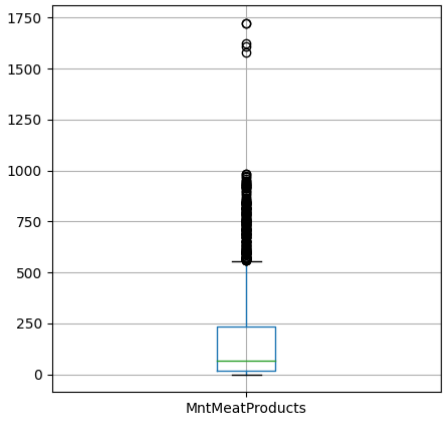
نمودار ۱۰- نمودار box plot مربوط به ستون MntFishProducts



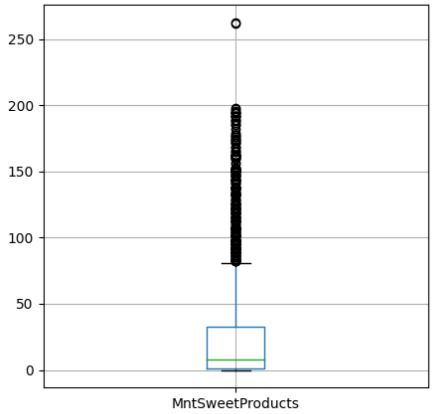
نمودار ۱۱- نمودار box plot مربوط به ستون MntFruits



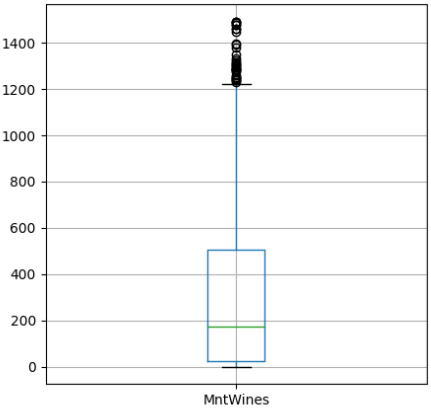
نمودار ۱۲- نمودار box plot مربوط به ستون MntGoldProds



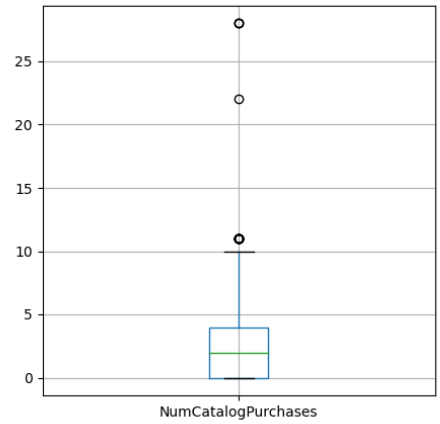
نمودار ۱۳- نمودار box plot مربوط به ستون MntMeatProducts



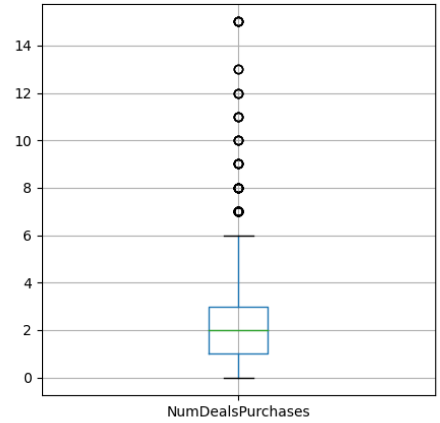
نمودار ۱۴- نمودار box plot مربوط به ستون MntSweetProducts



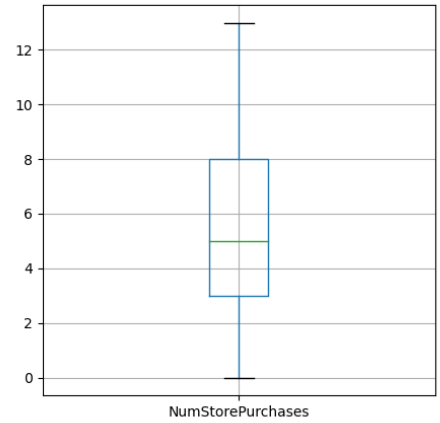
نمودار ۱۵- نمودار box plot مربوط به ستون MntWines



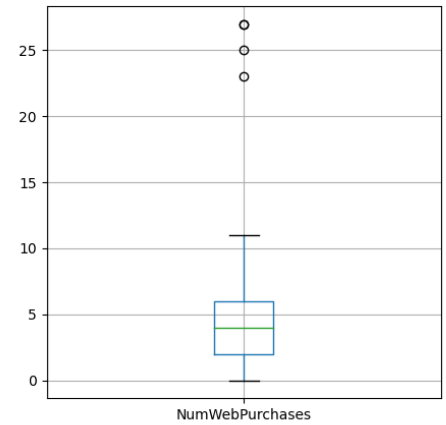
نمودار ۱۶- نمودار box plot مربوط به ستون NumCatalogPurchases



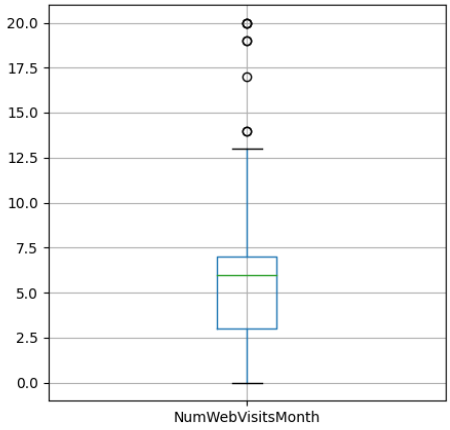
نمودار ۱۷- نمودار box plot مربوط به ستون NumDealsPurchases



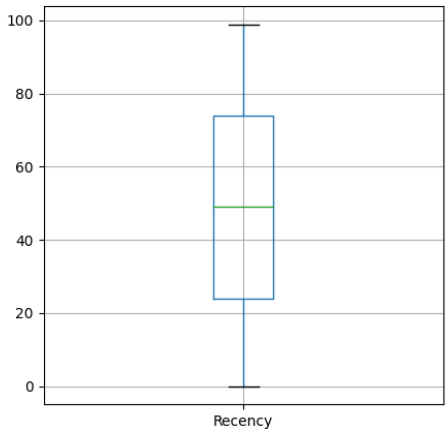
نمودار ۱۸- نمودار box plot مربوط به ستون NumStorePurchases



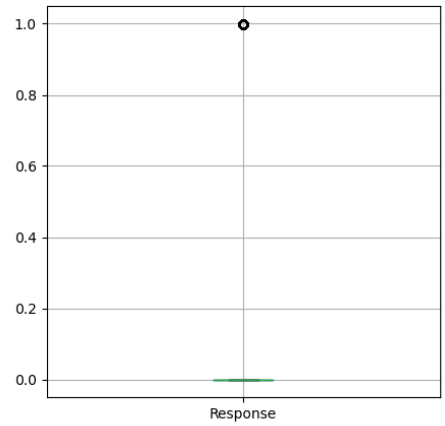
نمودار ۱۹- نمودار box plot مربوط به ستون NumWebPurchases



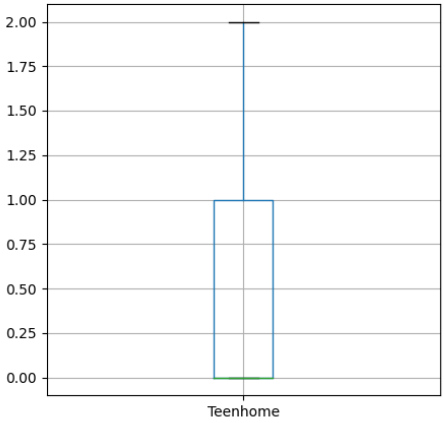
نمودار ۲۰- نمودار box plot مربوط به ستون NumWebVisitsMonth



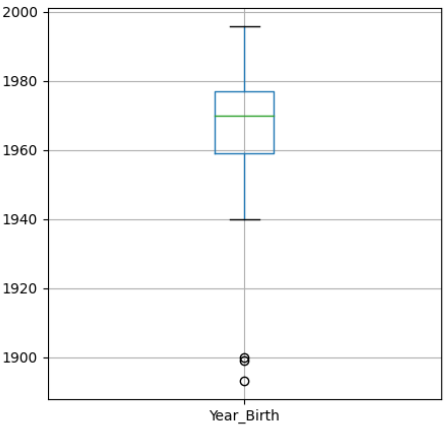
نمودار ۲۱- نمودار box plot مربوط به ستون Recency



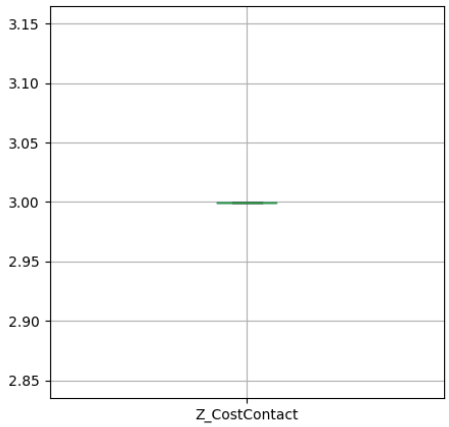
نمودار ۲۲- نمودار box plot مربوط به ستون Response



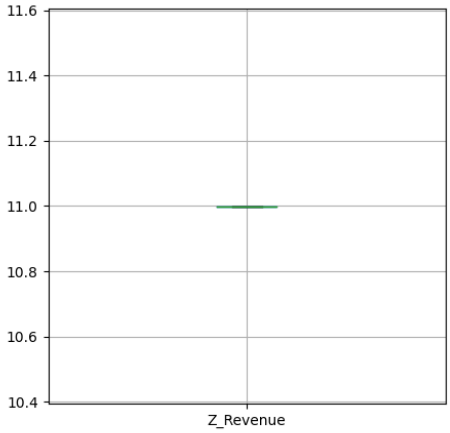
نمودار ۲۳- نمودار box plot مربوط به ستون Teenhome



نمودار ۲۴- نمودار box plot مربوط به ستون Year_Birth



نمودار ۲۵- نمودار box plot مربوط به ستون Z_CostContact



نمودار ۲۶- نمودار box plot مربوط به ستون Z_Revenue

۲-۱-۳- بخش سوم

در جدول زیر نام ستون ها به همراه نوع آن ها (Nominal, Ordinal, Binary, Numeric) آمده است:

جدول ۱- این جدول نوع هریک از ستون های دیتاست را نشان میدهد.

شماره	نام ستون	نوع ستون
۱	ID	Nominal
۲	Year_Birth	Nominal
۳	Education	Ordinal
۴	Marital_Status	Nominal
۵	Income	Numeric
۶	Kidhome	Numeric
۷	Teenhome	Numeric
۸	Dt_Customer	Nominal
۹	Recency	Numeric

۱۰	MntWines	Numeric
۱۱	MntFruits	Numeric
۱۲	MntMeatProducts	Numeric
۱۳	MntFishProducts	Numeric
۱۴	MntSweetProduct	Numeric
۱۵	MntGoldProds	Numeric
۱۶	NumDealsPurchases	Numeric
۱۷	NumWebPurchases	Numeric
۱۸	NumCatalogPurchases	Numeric
۱۹	NumStorePurchases	Numeric
۲۰	NumWebVisitsMonth	Numeric
۲۱	AcceptedCmp3	Binary
۲۲	AcceptedCmp4	Binary
۲۳	AcceptedCmp5	Binary
۲۴	AcceptedCmp1	Binary
۲۵	AcceptedCmp2	Binary
۲۶	Complain	Binary
۲۷	Z_CostContact	Numeric
۲۸	Z_Revenue	Numeric
۲۹	Response	Binary

۲-۱-۴- بخش چهارم

برای صحت داده ها ما باید موارد زیر را چک کنیم:

۱. مقادیر null یا از دست رفته را چک کنیم.
۲. نوع داده همان نوع مدنظر باشد.
۳. برای ستون هایی که باید مقادیر آنها منحصر به فرد باشد این مورد را چک میکنیم.
۴. محدوده مقادیر همان محدوده مدنظر باشد.

ابتدا از لحاظ داده های از دست رفته و null ستون ها را بررسی میکنیم که برای اینکار از متد isnull() استفاده میکنیم و تعداد مقادیر null را بدست می آوریم:

```
marketing_campaign_df.isnull().sum()
```

```
ID                     0
Year_Birth             0
Education              0
Marital_Status         0
Income                24
Kidhome                0
Teenhome               0
Dt_Customer            0
Recency                0
MntWines               0
MntFruits              0
MntMeatProducts        0
MntFishProducts        0
MntSweetProducts       0
MntGoldProds           0
NumDealsPurchases      0
NumWebPurchases        0
NumCatalogPurchases    0
NumStorePurchases      0
NumWebVisitsMonth      0
AcceptedCmp3           0
AcceptedCmp4           0
AcceptedCmp5           0
AcceptedCmp1           0
AcceptedCmp2           0
Complain               0
Z_CostContact          0
Z_Revenue              0
Response               0
dtype: int64
```

تصویر ۱- تعداد مقادیر null به ازای هر ستون

همانطور که مشاهده میکنید تنها ستون income دارای تعداد کمی مقادیر null است.

پس از اینکار میخواهیم نوع داده ها، نوع مد نظر ما باشند، فلذا داده های عدد صحیح نوعشان را int32 و داده های اعشاری را float32 و داده های رشته ای object و در نهایت تاریخ را با متد to_datetime به نوع datetime تبدیل میکنیم.

```
ID                              int32
Year_Birth                      int32
Education                      object
Marital_Status                 object
Income                        float32
Kidhome                         int32
Teenhome                        int32
Dt_Customer           datetime64[ns]
Recency                         int32
MntWines                        int32
MntFruits                       int32
MntMeatProducts                 int32
MntFishProducts                 int32
MntSweetProducts                int32
MntGoldProds                    int32
NumDealsPurchases               int32
NumWebPurchases                 int32
NumCatalogPurchases             int32
NumStorePurchases               int32
NumWebVisitsMonth               int32
AcceptedCmp3                    int32
AcceptedCmp4                    int32
AcceptedCmp5                    int32
AcceptedCmp1                    int32
AcceptedCmp2                    int32
Complain                        int32
Z_CostContact                   int32
Z_Revenue                       int32
Response                        int32
dtype: object
```

تصویر ۲- نوع داده های هر ستون

ستون ID: باید مقادیر آن منحصر به فرد باشد که آن را چک با استفاده از صفت is_unique چک میکنیم.

ستون Year_Birth: مقادیر min و max آن را چک میکنیم تا در بازه درستی باشد که بازه (۱۸۹۳, ۱۹۹۶) را شامل میشود که صحیح است.

ستون Education: مقادیری که برای سطح تحصیلات در نظر گرفته شده را چک میکنیم تا معتبر باشند که این مقادیر به صورت زیر و معتبر هستند:

['Graduation', 'PhD', 'Master', 'Basic', '2n Cycle']

ستون Marital_Status: برای این ستون نیز ابتدا مقادیر را چک میکنیم، که به صورت زیر است:

['Single', 'Together', 'Married', 'Divorced', 'Widow', 'Alone', 'Absurd', 'YOLO']

تعداد تکرار هر مقدار به صورت زیر است:

Married 864
Together 580
Single 480
Divorced 232
Widow 77
Alone 3
Absurd 2
YOLO 2

همانطور که میبینیم مقادیر alone و absurd و YOLO که مقادیر نامرتبط و کمتری هستند را حذف میکنیم.

ستون Income: دیدیم که این ستون چند تا مقدار null دارد. چون تعداد این مقادیر کم است لذا این مقادیر را با میانگین آن ستون پر میکنیم.

در کل ستون های numeric را برای null نبودن مقادیر آنها چک میکنیم و تنها ستون عددی که مقادیر null داشت ستون Income بود که مقادیر آن را با میانگین ستون پر کردیم.

ستون های binary نیز تنها باید شامل دو مقدار صفر و ۱ باشند. این مورد را نیز برای تمام ستون های binary چک میکنیم.

۲-۱-۵- بخش پنجم

در این بخش ما باید کامل بودن داده ها را چک کنیم. ما این را در بخش قبل انجام دادیم و دیدیم تنها ستون Income دارای ۲۴ مقدار null بود و چون این ستون یک ستون عددی و مربوط به درآمد است لذا از میانگین ستون برای پرکردن این مقادیر استفاده کردیم.

از مد نیز میتوان استفاده کرد یا مثلا داده ها را براساس مقادیر ویژگی ها خوشه بندی کنیم و مقادیر null را برابر با میانگین یا مد آن خوشه قرار بدهیم.

یا مثلا با استفاده از KNN مقدار نزدیکترین داده به آنها را در ستون Income قرار دهیم. این ها روش هایی است که میتوانیم برای پر کردن این ستون از آنها استفاده کنیم.

همچنین در این بخش برای حذف داده های پرت ابتدا ستون های عددی را در نظر میگیریم سپس IQR و تعریف کران بالا و پایین برای داده ها، داده هایی که خارج از کران باشند را حذف میکنیم.

۲-۲- تحلیل EDA

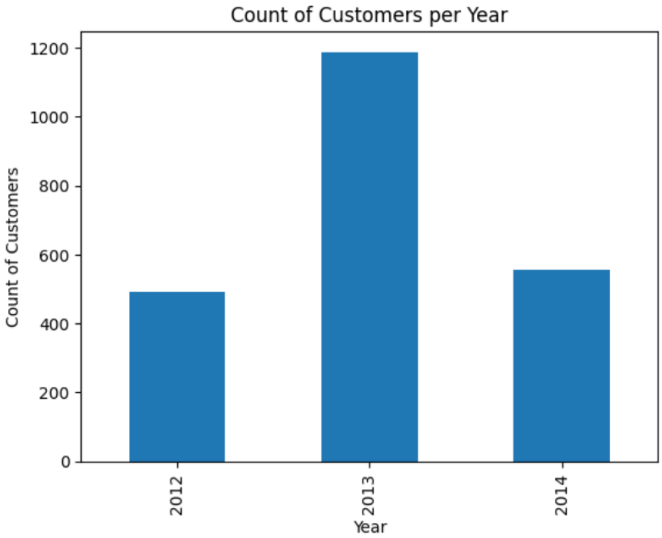
۲-۲-۱- بخش اول

با توجه به ابر کلمات برای تحصیلات مشتریان، مشاهده میکنیم که بیشتر مشتریان فارغ التحصیل می باشند و پس از آن مشتریان دارای تحصیلات دکترا و ارشد میباشند. در رتبه های آخر نیز تحصیلات سیکل و پایه قرار دارد.

تصویر ۳- ابرکلمات برای تحصیلات مشتریان

۲-۲-۲- بخش دوم

برای اینکه بفهمیم در چه سالی، بیشترین مشتریان در شرکت ثبتنام کرده است ما از نمودار هیستوگرام استفاده میکنیم. به این صورت که ابتدا بر اساس مقدار سال در ستون Dt_Customer سطر ها را گروه بندی میکنیم و تعداد سطر ها در هر گروه را با استفاده از متد ()count بدست می آوریم سپس نمودار هیستوگرام را براساس سال و تعداد مشتری ها در هر سال رسم میکنیم که به صورت زیر میشود:



نمودار ۲۷- تعداد مشتری هایی که در هر سال ثبتنام کرده اند.

همانطور که مشاهده میکنیم بیشترین تعداد مشتری در سال ۲۰۱۳ در شرکت ثبتنام کرده اند.

۲-۲-۳- بخش سوم

در این بخش ما میخواهیم نشان دهیم که در دو سال گذشته، مردم چه محصولاتی را بیشتر خریداری کرده اند. برای اینکار مقدار هر ستون مربوط به محصولات مثل گوشت، نوشیدنی، طلا، ماهی و ... را به صورت جداگانه جمع میزنیم و سپس مقادیر را با هم مقایسه میکنیم.

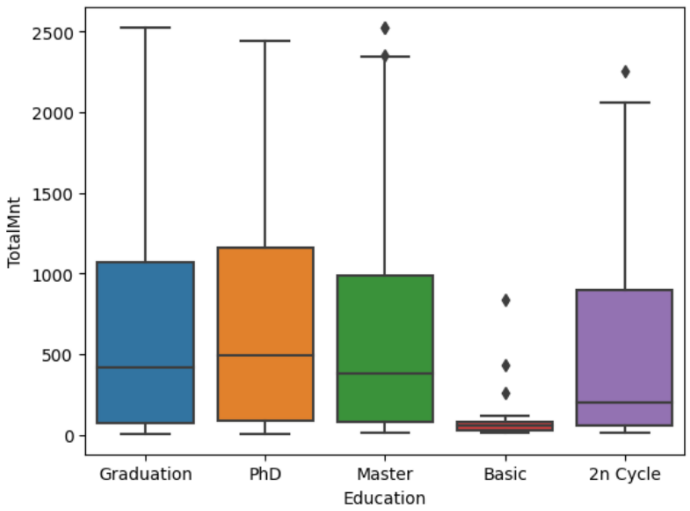
total_wine: 678907
total_fruit: 58730
total_meat: 373164
total_fish: 83615
total_sweet: 60533
total_gold:98036

با مقایسه اعداد متوجه میشویم که مردم از بین محصولات موجود، نوشیدنی را از همه بیشتر خریداری کرده اند. پس از آن گوشت و سپس طلا و بعد از آن ماهی و بعد شیرینی و در نهایت نیز کمترین میزان خرید متعلق به میوه است.

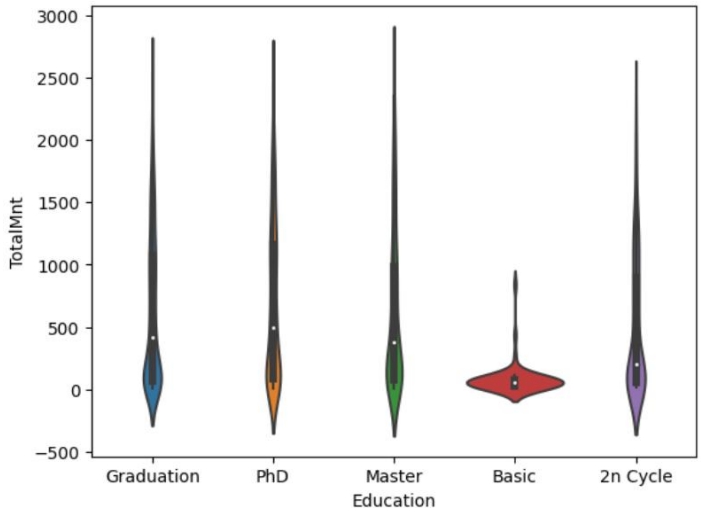
۲-۲-۴- بخش چهارم

در این بخش ما ابتدا مقادیر ستون های مرتبط با مبلغ صرف شده در دوسال گذشته را با هم جمع و در ستون جدیدی به نام TotalMnt قرار میدهیم.

حال برای فهمیدن تاثیر تحصیلات بر میلغ مصرف شده چون مبلغ مصرف شده یک مقدار numeric و تحصیلات یک مقدار ordinal است، پس ما از نمودار هایی مثل scatter برای فهمیدن ارتباط و همبستگی بین مقادیر این دو ستون نمیتوانیم استفاده کنیم. بنابراین برای تحلیل این دو ستون و تحلیل تاثیر یکی بر دیگری ما از نمودار box plot و violin plot استفاده میکنیم.



نمودار ۲۸- نمودار box plot برای تاثیر تحصیلات بر مبلغ مصرف شده در ۲ سال گذشته

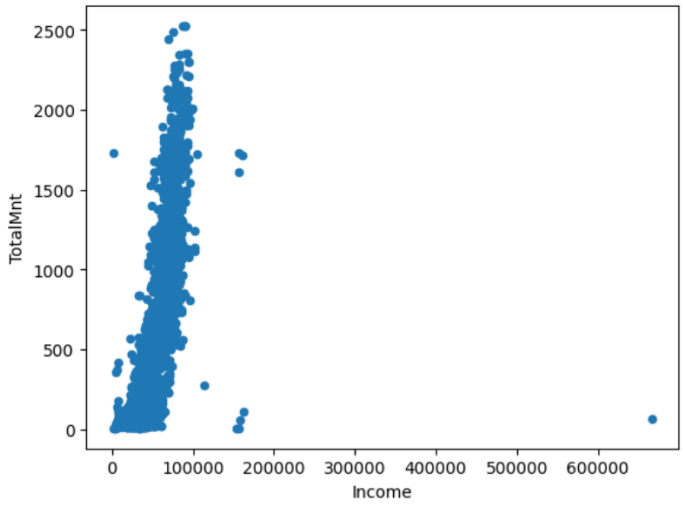


نمودار ۲۹- نمودار violin plot برای تاثیر تحصیلات بر مبلغ مصرف شده در ۲ سال گذشته

با تحلیل نمودار های بالا میفهمیم که مثلا کسانی که تحصیلات پایه دارند کمترین میزان خرید را دارند و میبینیم که بیشترین میزان خرید این افراد در بازه حدودا زیر ۱۰۰ است. از طرف دیگر هرچه سطح تحصیلات بیشتر میشود، تقریبا خرید افراد نیز بیشتر میشود. برای افراد فارغ التحصیل و ارشد و دکترا، میزان تراکم و میزان خرید تقریبا مشابه است. برای افرادی که سیکل دارند اما میزان خرید کمتر از افراد با تحصیلات بالاتر است و میبینیم که بیشتر حجم خرید های آنها زیر ۲۰۰ می باشد. تقریبا میتوان نتیجه گرفت که هرچه تحصیلات بیشتر باشد، افراد خرید بیشتری انجام داده اند.

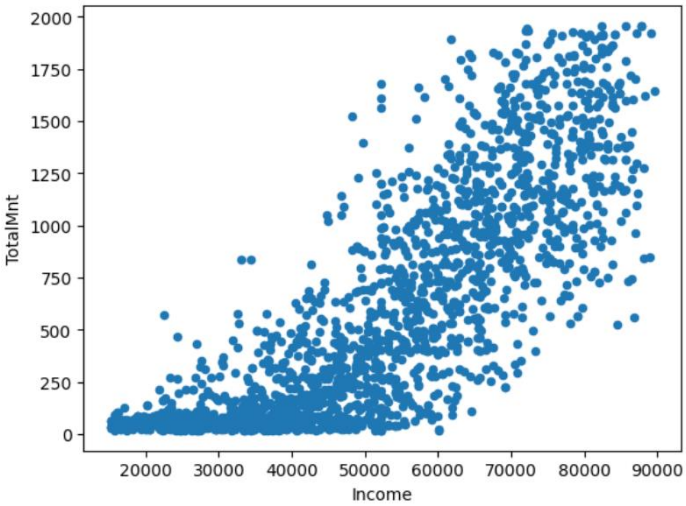
۲-۲-۵- بخش پنجم

برای بررسی همبستگی بین درآمد و مبلغ صرف شده هر فرد ما از نمودار scatter استفاده میکنیم. نمودار scatter اولی که روی کل داده ها رسم کردیم به صورت زیر است:



نمودار ۳۰- نمودار scatter برای درآمد و مبلغ صرف شده هر فرد روی کل داده ها

مشکل این نمودار این است که برخی داده های پرت در بین مقادیر درآمد وجود دارد درنتیجه نمودار به خوبی همبستگی را نشان نمیدهد، بنابراین در تلاش بعدی ما در رسم نمودار scatter از داده های outlier صرف نظر کردیم و نمودار به صورت زیر شد:

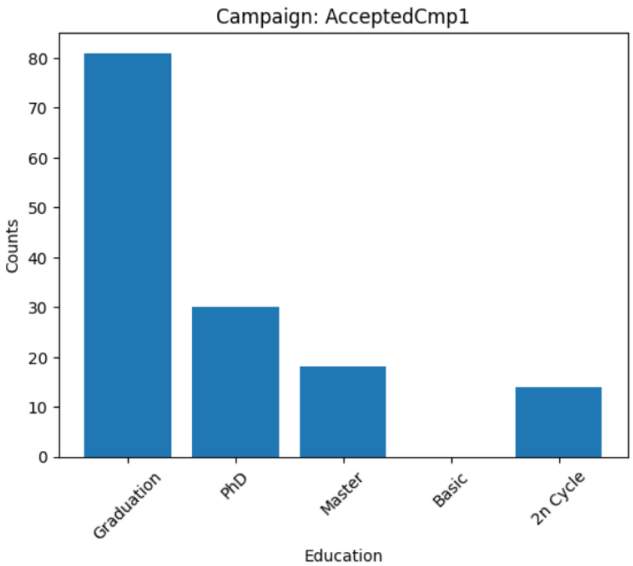


نمودار ۳۱- نمودار scatter برای درآمد و مبلغ صرف شده هر فرد صرف نظر از داده های پرت

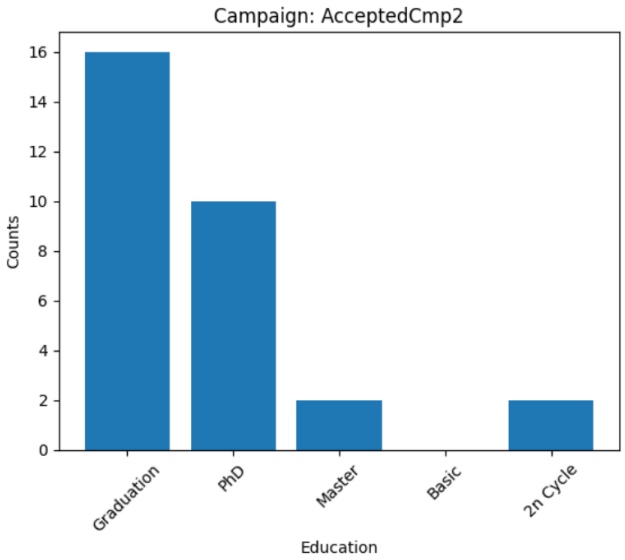
همانطور که مشاهده میکنیم، بین این دو همبستگی مثبت وجود دارد یعنی با افزایش درآمد، میزان خرید افراد نیز افزایش میابد. برای اطمینان ما مقدار همبستگی بین این دوستون را نیز بدست آوردیم که برابر با مقدار ۰.۶۶ می باشد که نشان دهنده همبستگی بین مقادیر این دو ستون می باشد.

۲-۲-۶- بخش ششم

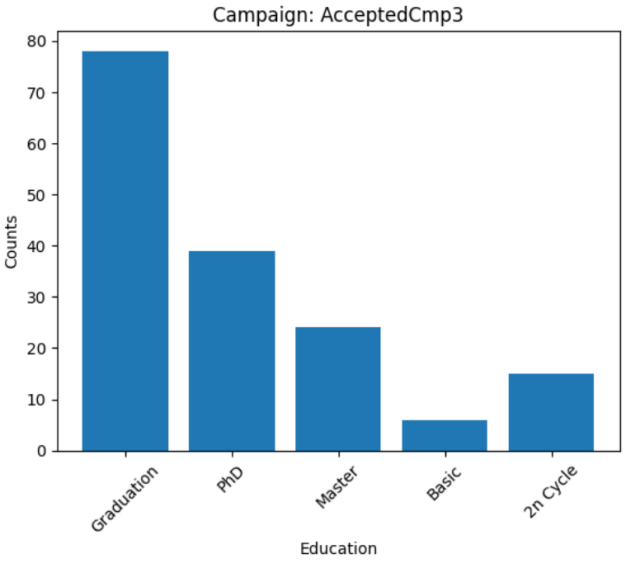
برای اینکه بفهمیم افراد با چه تحصیلاتی در هر کمپین شرکت کرده اند، ما یک دیکشنری برای مقاطع مختلف تحصیلی تعریف میکنیم و برای هر کمپین تعداد افراد در مقاطع مختلف را بدست می آوریم و نمودار هیستوگرام آن را به صورت زیر رسم میکنیم:



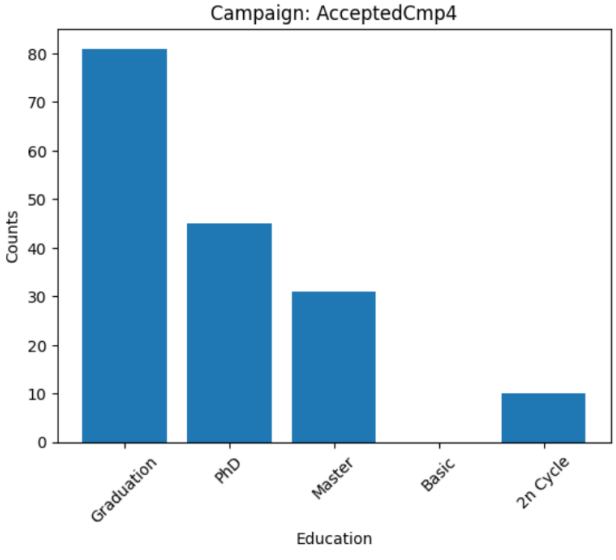
نمودار ۳۲- تعداد افرادی که در کمپین ۱ شرکت کردند بر حسب سطح تحصیلات نمایش داده شده است.



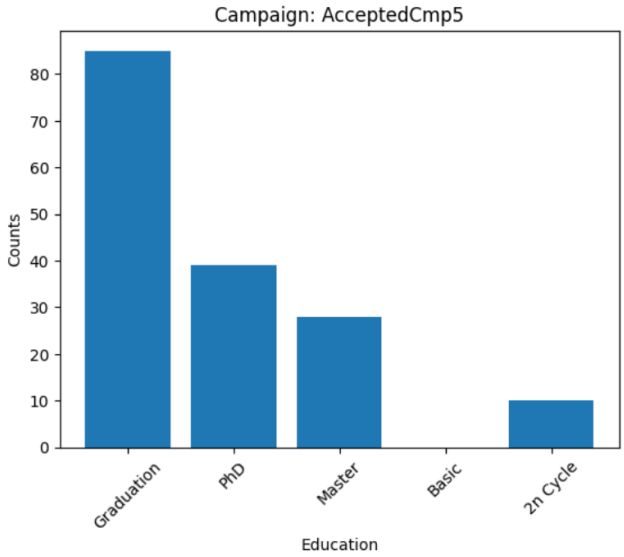
نمودار ۳۳- تعداد افرادی که در کمپین ۲ شرکت کردند بر حسب سطح تحصیلات نمایش داده شده است.



نمودار ۳۴- تعداد افرادی که در کمپین ۳ شرکت کردند بر حسب سطح تحصیلات نمایش داده شده است.



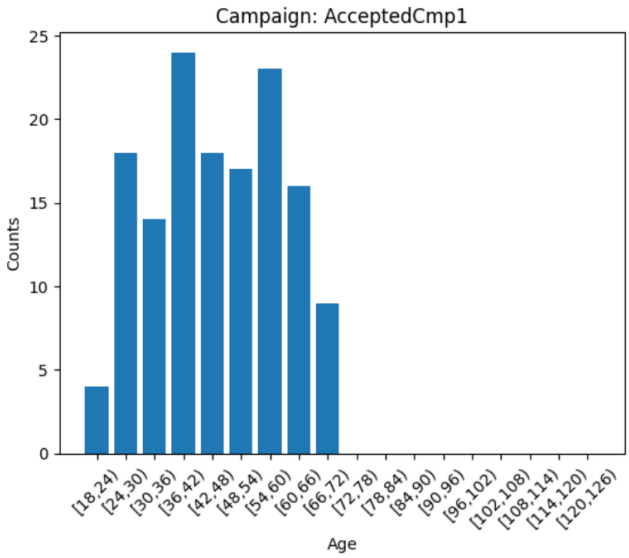
نمودار ۳۵- تعداد افرادی که در کمپین ۴ شرکت کردند بر حسب سطح تحصیلات نمایش داده شده است.



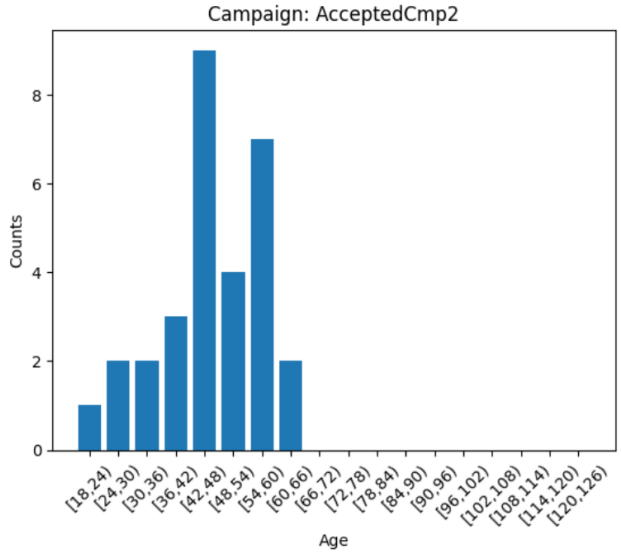
نمودار ۳۶- تعداد افرادی که در کمپین ۵ شرکت کردند بر حسب سطح تحصیلات نمایش داده شده است.

برای سن نیز ابتدا ما یک ستون جدید به نام Age ایجاد میکنیم که تفاضل سال ۲۰۱۴ (سال آن زمان) و سال تولد افراد که سن افراد را بدست می دهد می باشد.

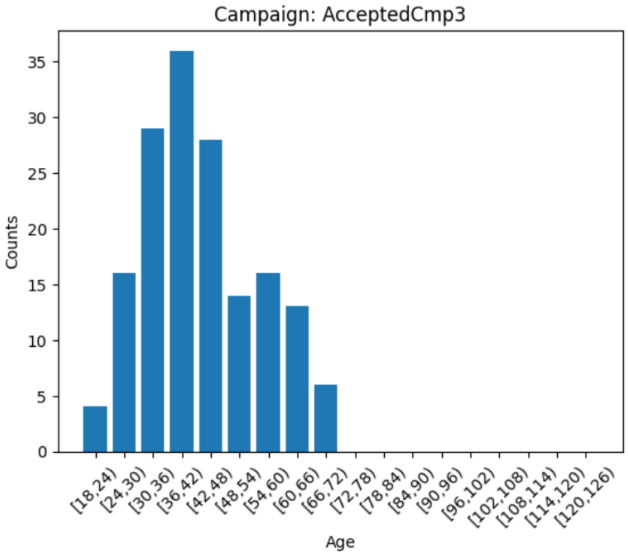
پس از اینکار ما bin های با سایز ۶ در نظر میگیریم. کمترین سن ۱۸ است. ما روی بازه های مختلف سنی حرکت میکنیم و بعد تعداد افرادی که در آن کمپین شرکت کرده اند و در آن محدوده سنی هستند را بدست می آوریم و در نهایت برای هر کمپین نمودار هیستوگرام آن را رسم میکنیم که به صورت زیر می باشد:



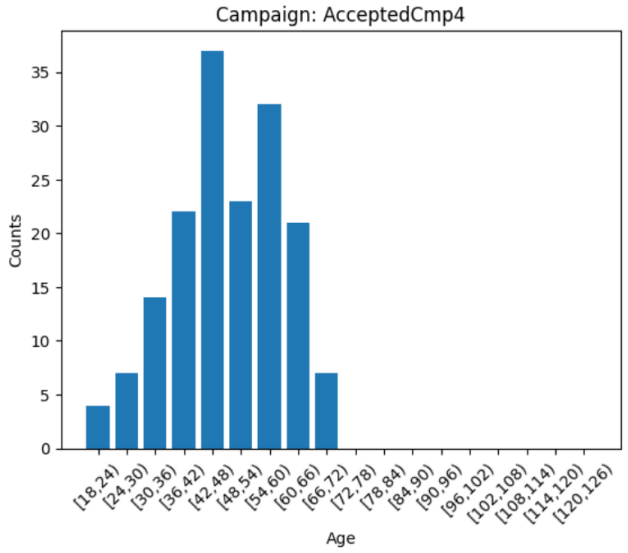
نمودار ۳۷- تعداد افرادی که در کمپین ۱ شرکت کردند بر حسب سن نمایش داده شده است.



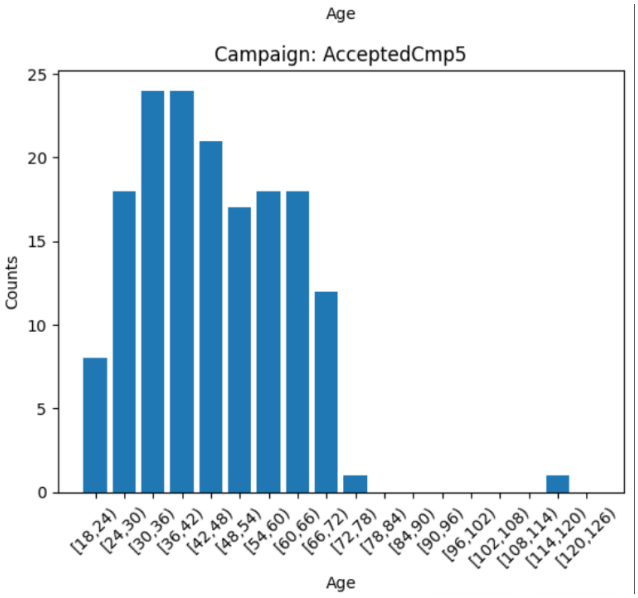
نمودار ۳۸- تعداد افرادی که در کمپین ۲ شرکت کردند بر حسب سن نمایش داده شده است.



نمودار ۳۹- تعداد افرادی که در کمپین ۳ شرکت کردند بر حسب سن نمایش داده شده است.



نمودار ۴۰- تعداد افرادی که در کمپین ۴ شرکت کردند بر حسب سن نمایش داده شده است.



نمودار ۴۱- تعداد افرادی که در کمپین ۵ شرکت کردند بر حسب سن نمایش داده شده است.

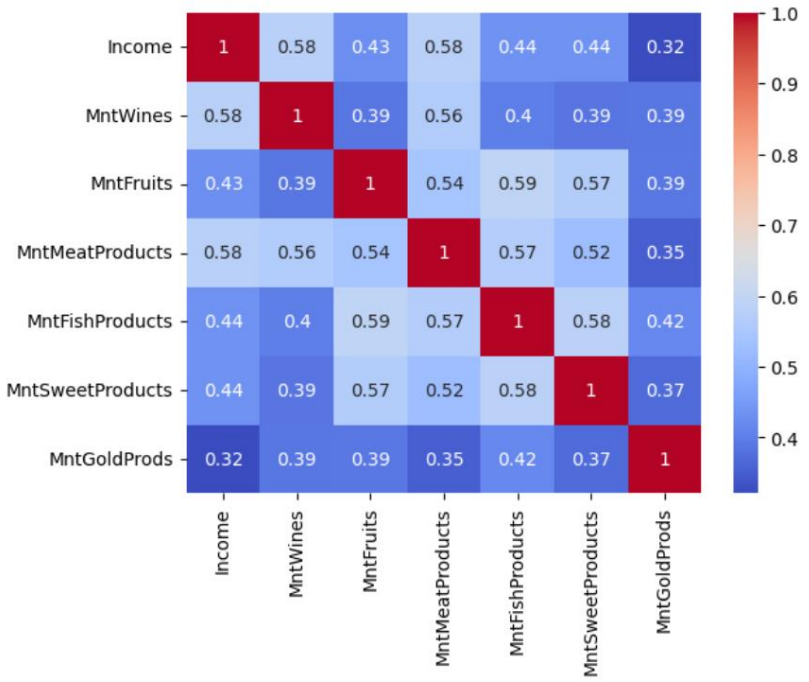
با تحلیل نمودار های بالا میتوانیم این نتیجه را بگیریم که مثلا در کمپین ۱ و ۵ اکثر رده های سنی حضور دارند، اما مثلا در کمپین ۲ و ۳ و ۴ اکثر افراد میان سال به بالا هستند.

۲-۳- تسک های نمره اضافه

۲-۳-۱- بخش اول

برای ارتباط سطح درآمد با مبلغ مصرف شده برای نوشیدنی، میوه، گوشت، ماهی، شیرینی و طلا ، ما از تابع corr در کتابخانه pandas برای ستون های نوشیدنی، میوه، گوشت، ماهی، شیرینی و طلا و ستون income استفاده میکنیم و مقدار correlation بین دو به دوی ستون ها را در قالب یک ماتریس ۲ در ۲ دریافت میکنیم که مقدار ۱ بیشترین میزان همبستگی و صفر کمترین میزان همبستگی را مشخص میکند.

پس از اینکار با استفاده از نمودار heatmap ماتریس همبستگی بدست آمده را نمایش میدهیم:



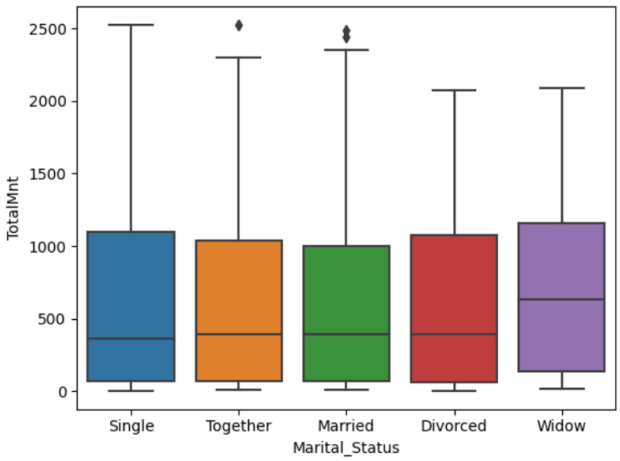
نمودار ۴۲- نمودار heatmap مربوط به همبستگی بین درآمد با نوشیدنی، میوه، گوشت، ماهی، شیرینی و طلا

سطر و ستون اول، همبستگی ستون های مدنظر با درآمد افراد را نشان میدهد. همانطور که مشاهده میکنید بیشترین همبستگی با درآمد مربوط به نوشیدنی و گوشت می باشد و کمترین همبستگی و ارتباط با درآمد مربوط به طلا می باشد.

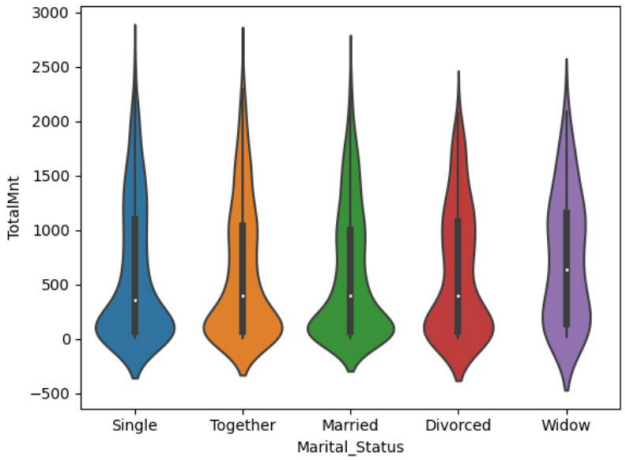
۲-۳-۲- بخش دوم

سوال ۱: درآمد و میزان خرید کلی افراد را براساس وضعیت تاهل را بررسی کنید.

برای این سوال مشابه سوال چهارم تحلیل EDA از نمودار box plot و violin plot استفاده میکنیم که به صورت زیر می باشد (محور افقی وضعیت های مختلف تاهل و محور عمودی مقادیر مبلغ صرف شده کل را نشان میدهد):



نمودار ۴۳- نمودار box plot برای تاثیر وضعیت تاهل بر مبلغ مصرف شده در ۲ سال گذشته



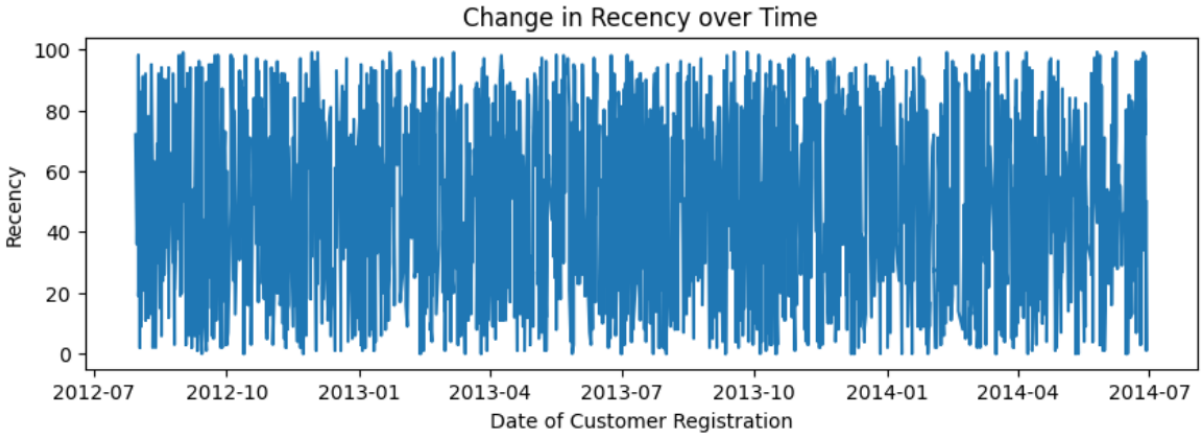
نمودار ۴۴- نمودار violin plot برای تاثیر وضعیت تاهل بر مبلغ مصرف شده در ۲ سال گذشته

از دو نمودار بالا میتوانیم بفهمیم که بیشترین مبلغ صرف شده مربوط به افراد سینگل :) است. تقریبا مشاهده میکنیم که بیشترین تراکم خرید برای تمام وضعیت های تاهل در بازه بین حدود ۱۰۰ تا ۴۰۰ می باشد.

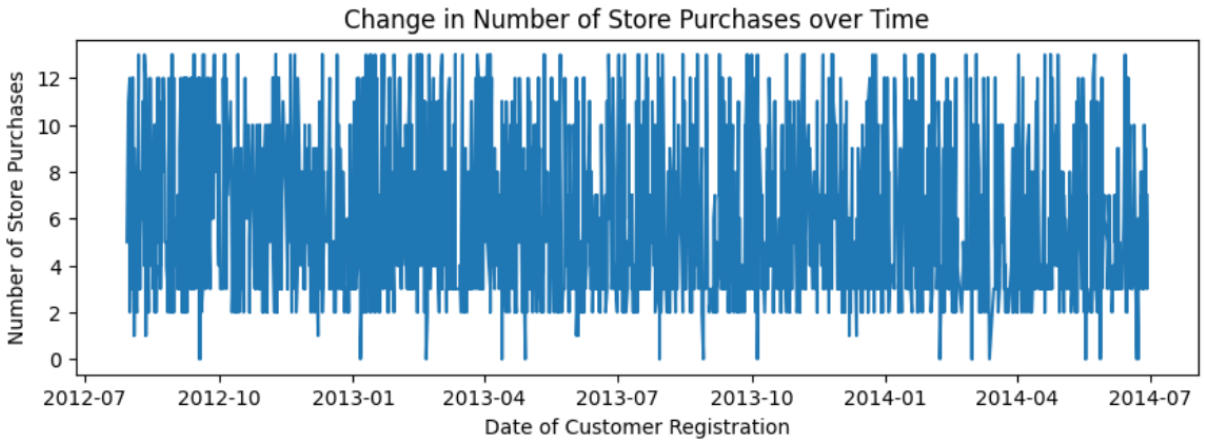
در تمام این وضعیت ها مقدار مینیمم تقریبا مشابه است. پراکندگی نیز تا حدودی بین Single و Together و Married و Divorced مشابه است و فقط بریا حالت widow کمی نمودار متفاوت است و میزان پراکندگی فرق میکند.

سوال ۲: با توجه به ستون Dt_Customer که تاریخ ثبت مشتری را نشان می‌دهد، بررسی روند تغییر Recency یا تعداد خریدهای مستقیم از فروشگاه بر اساس زمان ثبت مشتری مورد بررسی قرار دهید.

Recency تعداد روز های پس از آخرین خرید را نشان میدهد و NumStorePurchases تعداد خرید هایی است که مستقیما از فروشگاه ها انجام شده است. ما برای این بررسی ابتدا داده های دیتاست را براساس ستون Dt_Customer مرتب میکنیم و سپس با نمودار های مناسب این دو را در طول زمان مورد بررسی قرار میدهیم:

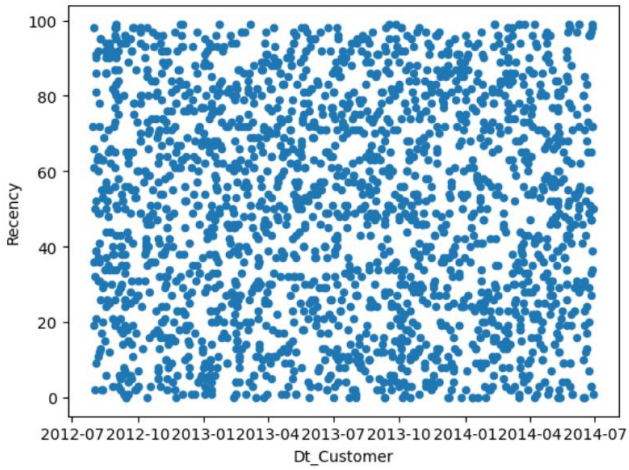


نمودار ۴۵- نمودار خطی تعداد روز های پس از آخرین خرید بر جسب زمان

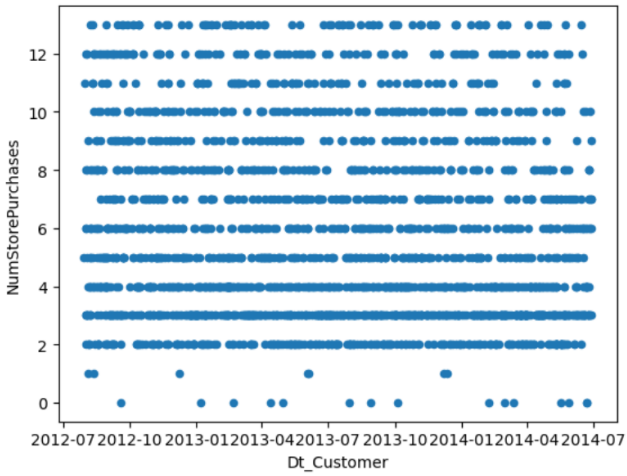


نمودار ۴۶- نمودار خطی تعداد خرید های مستقیم از فروشگاه ها بر حسب زمان

از این دو نمودار الگوی خاصی قابل استخراج نیست، پس ما نمودار scatter را برای این دو رسم میکنیم:



نمودار ۴۷- نمودار scatter تعداد روز های پس از آخرین خرید و زمان



نمودار ۴۸- نمودار scatter تعداد خرید های مستقیم از فروشگاه ها و زمان

همانطور که مشاهده میکنیم بین تغییرات زمانی و تعداد روز های پس از آخرین خرید یا تعداد خرید های مستقیم از فروشگاه هیچگونه ارتباط و همبستگی وجود ندارد. لذا الگوی منظمی بین زمان و تعداد روز های بعد از آخرین خرید یا تعداد خرید های مستقیم از فروشگاه نیست و نمیتوان دانش خاصی را استخراج کرد. البته این مورد طبیعی است چون افراد هر زمان که بخواهند

میتوانند حضوری یا اینترنتی خرید کنند یا هرموقع که نیاز داشته باشند خرید ها را انجام دهند و این از الگوی خاصی تبعیت نمیکند و افراد با هم تفاوت دارند.

۳- فاز دوم پروژه

۳-۱- تحلیل الگو و ارائه مدل های پیش بینی

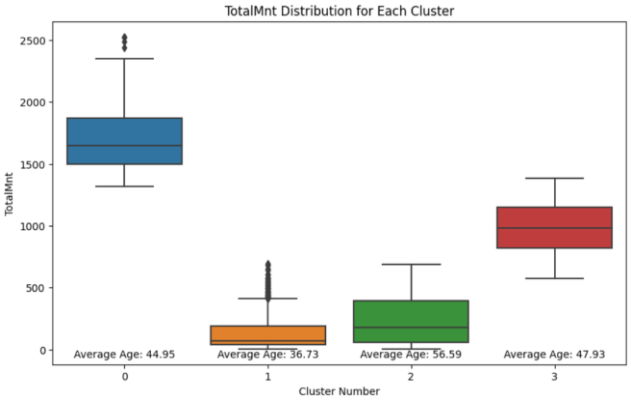
۳-۱-۱- بخش اول

در شروع باید ستون های مرتبط را پیدا کنیم. اول از همه ستون های زیر را به عنوان ستون های مرتبط در نظر گرفتم:

```
relevant_columns = ['Age', 'MntWines', 'MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds', 'TotalMnt']
```

اما Silhouette coefficient برای خوشه بندی مدنظر عدد ۰.۳۸ بود که نشان میدهد خوشه ها همپوشانی زیادی دارند. بعد از این تنها ستون های Age و TotalMnt را در نظر گرفتم و همین باعث شده که مقدار این ضریب به ۰.۴۱ تغییر پیدا کند.

خوشه بندی را روی ستون های مرتبط (یعنی TotalMnt و Age) انجام دادیم. سپس برای هر خوشه بر اساس ستون TotalMnt نمودار boxplot را رسم میکنیم و نمودار را براساس آن تحلیل میکنیم.



نمودار ۴۹- نمودار boxplot مربوط به میزان پراکندگی کل مبلغ مصرف شده در هر خوشه

همانطور که مشاهده میکنیم میانگین سنی هر خوشه زیر نمودار مربوط به آن آورده شده است. با توجه به نمودار دو خوشه ی ۰و۳ خوشه های rich محسوب میشوند و دو خوشه ی ۱و۲ خوشه های poor می باشند. از طرفی با توجه به میانگین سنی هر خوشه، دو خوشه ی ۰و۱ را را به عنوان young و دو خوشه ی دیگر را poor در نظر میگیریم.

لذا با توجه به این تحلیل خوشه ها را به صورت زیر در نظر میگیریم:

خوشه صفر: young-rich

خوشه یک: young-poor

خوشه دو: old-poor

خوشه سه: old-rich

همچنین خوشه بندی را با توجه به اینکه داده لیبل دار نداشتیم از روش Intrinsic برای ارزیابی کیفیت خوشه ها استفاده کردیم. برای اینکار از معیار Silhouette coefficient استفاده کردیم که مقادیر آن بین بازه -۱ و ۱ می باشد:

۱. مقدار نزدیک به ۱ نشان می دهد که نمونه ها به خوبی خوشه بندی شده اند و از خوشه های همسایه دور هستند.

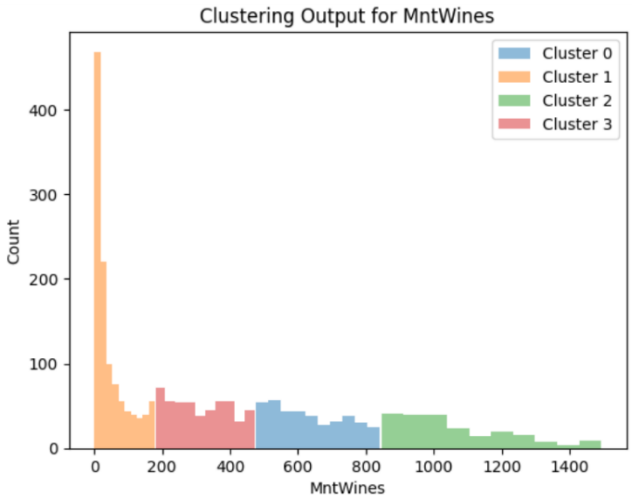
۲. مقدار نزدیک به ۰ نشان می دهد که خوشه های همپوشانی دارند یا اینکه نمونه ها روی مرز تصمیم بین خوشه ها یا بسیار نزدیک به آن هستند.

۳. مقدار نزدیک به -۱ نشان می دهد که نمونه ها به اشتباه خوشه بندی شده اند یا به خوشه اشتباهی اختصاص داده شده اند.

با توجه به اینکه مقدار این معیار ۰.۴۱ شد نتیجه میگیریم که خوشه ها تا حدودی همپوشانی دارند و یا خیلی نزدیک به هم می باشند.

۳-۱-۲- بخش دوم

برای این بخش ما ابتدا ستون هایی که میخواهیم به یکی از دسته های Non، Low، Frequent و High خوشه بندی کنیم را در نظر میگیریم. سپس برای هر ستون به صورت جداگانه بر اساس آن ستون ۴ خوشه درست میکنیم و نمودار مربوط به آن خوشه ها را رسم میکنیم تا آنها را تحلیل کنیم:



نمودار ۵۰- نمودار هسیتوگرام مربوط به خوشه بندی میزان مصرف نوشیدنی

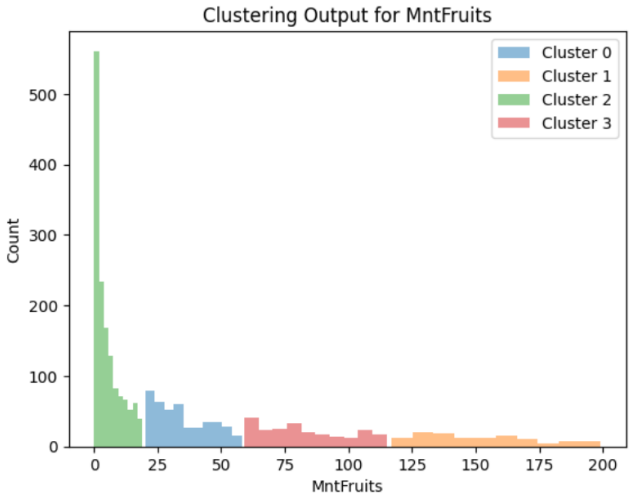
همانطور که مشاهده میکنیم هر رنگ نشان دهنده یکی از خوشه ها می باشد. براساس میزان مصرف نوشیدنی این خوشه ها را به صورت زیر نامگذاری میکنیم:

خوشه نارنجی: High_wines

خوشه قرمز: Frequent_wines

خوشه آبی: Low_wines

خوشه سبز: Non_wines



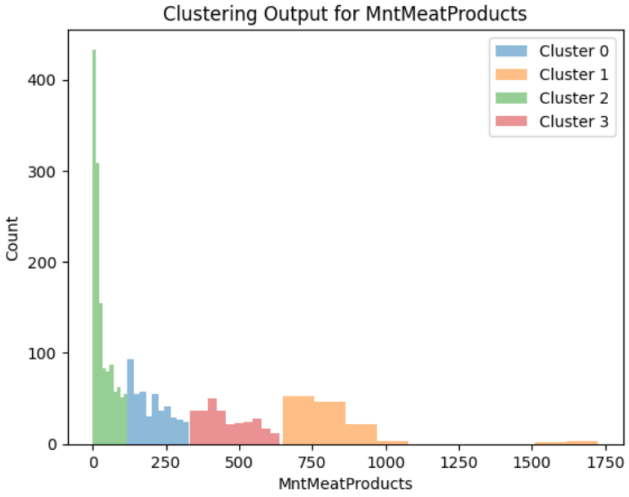
نمودار ۵۱- نمودار هیستوگرام مربوط به خوشه بندی میزان مصرف میوه

خوشه سبز: High_fruits

خوشه آبی: Frequent_fruits

خوشه قرمز: Low_fruits

خوشه نارنجی: Non_fruits



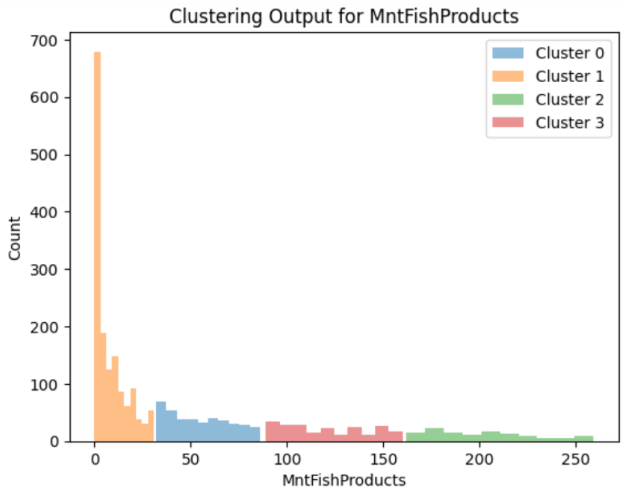
نمودار ۵۲- نمودار هیستوگرام مربوط به خوشه بندی میزان مصرف گوشت

خوشه سبز: High_meat

خوشه آبی: Frequent_meat

خوشه قرمز: Non_meat

خوشه نارنجی: Low_meat



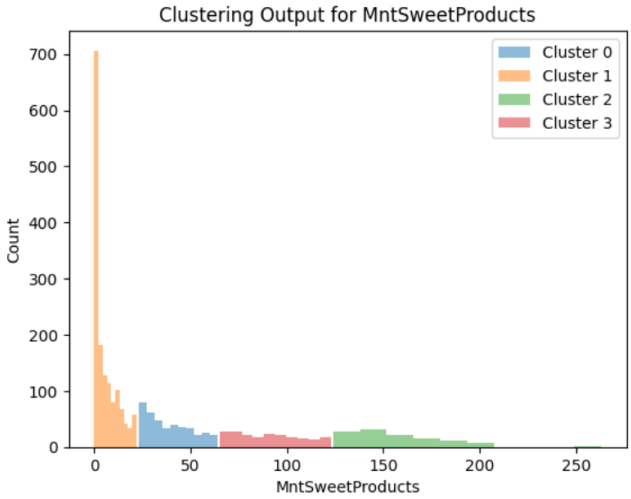
نمودار ۵۳- نمودار هیستوگرام مربوط به خوشه بندی میزان مصرف ماهی

خوشه نارنجی: High_fish

خوشه آبی: Frequent_fish

خوشه قرمز: Low_fish

خوشه سبز: Non_fish



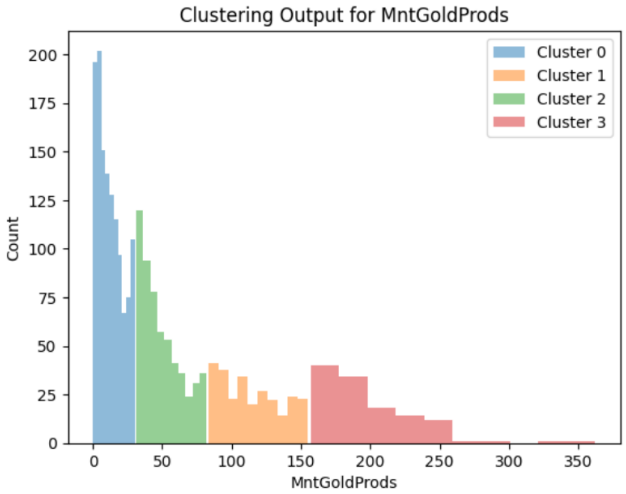
نمودار ۵۴- نمودار هیستوگرام مربوط به خوشه بندی میزان مصرف شیرینی

خوشه نارنجی: High_sweet

خوشه آبی: Frequent_sweet

خوشه قرمز: Non_sweet

خوشه سبز: Low_sweet



نمودار ۵۵- نمودار هیستوگرام مربوط به خوشه بندی میزان مصرف طلا

خوشه آبی: High_gold

خوشه سبز: Frequent_gold

خوشه نارنجی: Non_gold

خوشه قرمز: Low_gold

همچنین بر اساس معیار Silhouette coefficient کیفیت خوشه بندی ها را ارزیابی کردیم که نتیجه ارزیابی در جدول زیر آورده شده است:

جدول ۲- کیفیت خوشه بندی با معیار Silhouette Coefficient

شماره	خوشه بندی روی ستون	Silhouette coefficient
۱	MntWines	0.64
۲	MntFruits	0.68
۳	MntMeatProducts	0.67
۴	MntFishProducts	0.69
۵	MntSweetProducts	0.69
۶	MntGoldProds	0.61

همانطور که مشاهده میکنیم مقدار ضرایب تقریبا نزدیک به ۱ می باشد که نشان میدهد تا حدود خوبی خوشه ها از هم متمایز شده اند.

۳-۱-۳- بخش سوم

برای این بخش ابتدا ما ستون های مرتبط یعنی ستون های AcceptedCmp1وAcceptedCmp2 وAcceptedCmp3 و AcceptedCmp4وAcceptedCmp5 را در نظر گرفتیم. سپس یک تابع find_frequent_patterns تعریف کردیم که itemset های پرتکرار را محاسبه میکند.

ما برای مقادیر ۰.۰۲ و ۰.۰۵ و ۰.۲ و ۰.۴ و ۰.۵ و ۰.۶ برای min_support و min_confidence در نظر گرفتیم و برای مقادیر خیلی کم min_support و min_confidence الگوهایی را پیدا کرد هرچند چون مقدار min_support و min_confidence خیلی کوچک هستند عملا الگوی زیر، الگوهای پرتکرار در دیتاست محسوب نمیشوند.

جدول ۳- الگوهای پرتکرار با min_support=0.02 و min_confidence=0.02

itemsets	support	#
(AcceptedCmp1)	0.06	1
(AcceptedCmp3)	0.07	2
(AcceptedCmp4)	0.07	3
(AcceptedCmp5)	0.07	4
(AcceptedCmp1, AcceptedCmp4)	0.02	5
(AcceptedCmp1, AcceptedCmp5)	0.03	6
(AcceptedCmp5, AcceptedCmp4)	0.02	7

جدول ۴- الگوهای پرتکرار با min_support=0.02 و min_confidence=0.05

itemsets	support	#
(AcceptedCmp1)	0.06	1
(AcceptedCmp3)	0.07	2
(AcceptedCmp4)	0.07	3
(AcceptedCmp5)	0.07	4

جدول ۵- الگوهای پرتکرار با min_support=0.05 و min_confidence=0.02

itemsets	support	#
(AcceptedCmp1)	0.06	1
(AcceptedCmp3)	0.07	2
(AcceptedCmp4)	0.07	3
(AcceptedCmp5)	0.07	4

جدول ۶- الگوهای پرتکرار با min_support=0.05 و min_confidence=0.05

itemsets	support	#
(AcceptedCmp1)	0.06	1
(AcceptedCmp3)	0.07	2
(AcceptedCmp4)	0.07	3
(AcceptedCmp5)	0.07	4

۳-۱-۴- بخش چهارم

برای پیش بینی اینکه آیا کاربر در کمپین های آینده شرکت میکند یا خیر، ستون Response را به عنوان برچسب داده ها و ستون های زیر را به عنوان ویژگی های مرتبط با این پیش بینی در نظر میگیریم:

['Income', 'Kidhome', 'Teenhome', 'MntWines', 'MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds', 'NumDealsPurchases', 'NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases', 'NumWebVisitsMonth', 'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5', 'AcceptedCmp1', 'AcceptedCmp2']

برای این پیش بینی از دو طبقه بند naïve bayes و decision tree و همچنین مدل آنسامبل random forest classifier استفاده کردیم.

ابتدا دو طبقه بند naïve bayes و decision tree را روی داده های غیر نرمال آموزش میدهیم.

همچنین ما در اینجا ۸۰ درصد داده ها را برای آموزش و ۲۰ درصد داده ها را برای تست نگه داشته ایم و برای ارزیابی مدل ها از معیار های Accuracy، Precision، Recall و F1-Score استفاده کرده ایم.

برای داده های غیر نرمال معیار های ارزیابی، مقادیر زیر را محاسبه کردند:

Bayes Classification Metrics:
Accuracy: 0.7293064876957495
Precision: 0.23853211009174313
Recall: 0.40625
F1-Score: 0.30057803468208094

Decision Tree Classification Metrics:
Accuracy: 0.8120805369127517
Precision: 0.3611111111111111
Recall: 0.40625
F1-Score: 0.38235294117647056

همانطور که مشاهده میکنیم، در حالتی که داده ها غیرنرمال هستند، درخت تصمیم بهتر از naïve Bayesian عمل کرده است که این به دلیل مزیت درخت تصمیم نسبت به روش های دیگر است که درخت تصمیم نیاز به پیش پردازش روی داده ها ندارد و میتوان داده ها را مستقیم برای آموزش مدل استفاده کرد.

وقتی داده ها را نرمال کردیم مجددا این دو مدل را آموزش دادیم که این دفعه معیار های ارزیابی، مقادیر زیر را در خروجی دادند:

Bayes Classification Metrics:
Accuracy: 0.8053691275167785
Precision: 0.367816091954023
Recall: 0.5
F1-Score: 0.423841059602649

Decision Tree Classification Metrics:
Accuracy: 0.7941834451901566
Precision: 0.32051282051282054
Recall: 0.390625
F1-Score: 0.35211267605633806

همانطور که مشاهده میکنیم وقتی داده ها را نرمال کردیم مدل naïve Bayesian بهتر از Decision Tree عمل کرده است. این عملکرد بهتر naïve Bayesian در مقایسه با decision tree ممکن است به این دلیل باشد که ویژگی‌های موجود در مجموعه داده مقیاس‌های متفاوتی دارند و نرمال سازی آن‌ها به Naive Bayes کمک کرده است تا از اطلاعات هر ویژگی بهتر استفاده کند و منجر به بهبود عملکرد شود.

همچنین درخت‌های تصمیم بر اساس آستانه ویژگی‌ها تقسیم می‌کنند و مرزهای تصمیم تحت تأثیر مقیاس یا توزیع ویژگی‌ها قرار نمی‌گیرند. بنابراین، نرمال سازی ممکن است تأثیر قابل توجهی بر عملکرد درختان تصمیم نداشته باشد.

پس از اینها ما از مدل ensemble برای دسته بندی داده ها استفاده کردیم که برای اینکار از random forest classifier استفاده کردیم. عملکرد این مدل روی داده های تست به صورت زیر می باشد:

Random Forest Classification Metrics:
Accuracy: 0.8769574944071589
Precision: 0.6551724137931034
Recall: 0.296875
F1-Score: 0.4086021505376344

همانطور که میبینیم روش ensemble نسبت به دو روش قبل از کیفیت و عملکرد بهتری برخوردار است. این به دلیل این است که ما در روش random forest classifier چندین درخت را با هم ترکیب و از نتایج آنها به طور توام استفاده میکنیم و این باعث بهبود عملکرد میشود. توجه به این نکته نیز حائز اهمیت است که این نتیجه بدون نرمال سازی روی داده ها می باشد از طرفی نرمال سازی داده ها تاثیری در بهبود دقت این مدل نداشت.

۳-۱-۵- بخش پنجم

برای این قسمت نیز مشابه بخش قبل از مدل های naïve Bayesian و Decision Tree و Random Forest استفاده کرده ایم و نتایج به صورت زیر می باشد:

برای داده های غیر نرمال:

Bayes Classification Metrics:
Accuracy: 0.8076062639821029
Precision: 0.3877551020408163
Recall: 0.59375
F1-Score: 0.46913580246913583

Decision Tree Classification Metrics:
Accuracy: 0.8769574944071589
Precision: 0.6451612903225806
Recall: 0.3125
F1-Score: 0.42105263157894735

Random Forest Classification Metrics:
Accuracy: 0.8814317673378076
Precision: 0.6774193548387096
Recall: 0.328125
F1-Score: 0.4421052631578948

همانطور که انتظار داشتیم مدل Random Forest از دو مدل دیگر بهتر عمل کرده. همچنین در این بخش همانطور که مشاهده میکنیم مدل Decision Tree خیلی بهتر از naïve Bayesian عمل کرده است که باز هم به دلیل مزیت این روش نسبت به روش naïve Bayesian می باشد.

برای داده های نرمال شده:

Bayes Classification Metrics:
Accuracy: 0.8076062639821029
Precision: 0.3877551020408163
Recall: 0.59375
F1-Score: 0.46913580246913583

Decision Tree Classification Metrics:

Accuracy: 0.8769574944071589
Precision: 0.6451612903225806
Recall: 0.3125
F1-Score: 0.42105263157894735

همانطور که مشاهده میکنیم نتیجه برای حالت داده های نرمال تغییر نکرده و این نشان میدهد که نرمال سازی داده ها تاثیری در بهبود دقت این دو مدل در این بخش نداشته است.

۳-۲- تسک های نمره اضافه

۳-۲-۱- بخش اول

در این بخش برای تاثیر سطح تحصیلات در نحوه خرید مشتری از فروشگاه (کاتالوگ، وب سایت، حضوری) ۴ تا ستون Educationو NumWebPurchasesو NumCatalogPurchasesو NumStorePurchases را مد نظر قرار میدهیم و برای تحلیل از قوانین انجمنی و ابرکلمات استفاده خواهیم کرد.

ابتدا برای این بخش الگوهای پرتکرار را پیدا میکنیم که این قسمت را در بخش دوم تسک های نمره اضافه انجام دادیم و میتوانید برای اطلاعات بیشتر از نحوه انجام این گام به بخش مربوطه مراجعه کنید.

پس از بدست آوردن الگوهای پرتکرار با تعیین min_confidence=0.5 قوانین انجمنی را بدست می آوریم که قوانین مرتبط با سطح تحصیلات و نحوه خرید از فروشگاه را در جدول زیر آورده ایم:

جدول ۷- قوانین انجمنی برای سطح تحصیلات و نحوه خرید از فروشگاه با min_support=0.3 و min_confidence=0.5

#	support	confidence	antecedents	consequents
1	0.49	0.98	('Education_Graduation',)	('NumWebPurchases',)
2	0.49	0.5	('NumWebPurchases',)	('Education_Graduation',)
3	0.37	0.74	('Education_Graduation',)	('NumCatalogPurchases',)
4	0.37	0.51	('NumCatalogPurchases',)	('Education_Graduation',)
5	0.5	0.5	('NumStorePurchases',)	('Education_Graduation',)
6	0.5	0.99	('Education_Graduation',)	('NumStorePurchases',)
7	0.37	0.99	('Education_Graduation', 'NumCatalogPurchases')	('NumWebPurchases',)
8	0.37	0.75	('Education_Graduation', 'NumWebPurchases')	('NumCatalogPurchases',)
9	0.37	0.51	('NumCatalogPurchases', 'NumWebPurchases')	('Education_Graduation',)
10	0.37	0.74	('Education_Graduation',)	('NumCatalogPurchases', 'NumWebPurchases')
11	0.37	0.5	('NumCatalogPurchases',)	('Education_Graduation', 'NumWebPurchases')
12	0.49	0.98	('NumStorePurchases', 'Education_Graduation')	('NumWebPurchases',)
13	0.49	0.5	('NumStorePurchases', 'NumWebPurchases')	('Education_Graduation',)
14	0.49	1	('Education_Graduation', 'NumWebPurchases')	('NumStorePurchases',)
15	0.49	0.97	('Education_Graduation',)	('NumStorePurchases', 'NumWebPurchases')

('NumStorePurchases', 'Education_Graduation')	('NumWebPurchases',)	0.5	0.49	*16*
('NumCatalogPurchases',)	('NumStorePurchases', 'Education_Graduation')	0.75	0.37	*17*
('Education_Graduation',)	('NumStorePurchases', 'NumCatalogPurchases')	0.51	0.37	*18*
('NumStorePurchases',)	('Education_Graduation', 'NumCatalogPurchases')	1	0.37	*19*
('NumStorePurchases', 'NumCatalogPurchases')	('Education_Graduation',)	0.74	0.37	*20*
('NumStorePurchases', 'Education_Graduation')	('NumCatalogPurchases',)	0.5	0.37	*21*
('NumWebPurchases',)	('NumStorePurchases', 'Education_Graduation', 'NumCatalogPurchases')	1	0.37	*22*
('NumCatalogPurchases',)	('NumStorePurchases', 'Education_Graduation', 'NumWebPurchases')	0.76	0.37	*23*
('Education_Graduation',)	('NumStorePurchases', 'NumCatalogPurchases', 'NumWebPurchases')	0.51	0.37	*24*
('NumStorePurchases',)	('Education_Graduation', 'NumCatalogPurchases', 'NumWebPurchases')	1	0.37	*25*
('NumCatalogPurchases', 'NumWebPurchases')	('NumStorePurchases', 'Education_Graduation')	0.74	0.37	*26*
('Education_Graduation', 'NumWebPurchases')	('NumStorePurchases', 'NumCatalogPurchases')	0.5	0.37	*27*
('NumStorePurchases', 'NumWebPurchases')	('Education_Graduation', 'NumCatalogPurchases')	0.99	0.37	*28*
('NumStorePurchases', 'NumCatalogPurchases')	('Education_Graduation', 'NumWebPurchases')	0.75	0.37	*29*
('NumStorePurchases', 'Education_Graduation')	('NumCatalogPurchases', 'NumWebPurchases')	0.51	0.37	*30*
('NumStorePurchases', 'NumCatalogPurchases', 'NumWebPurchases')	('Education_Graduation',)	0.74	0.37	*31*
('NumStorePurchases', 'Education_Graduation', 'NumWebPurchases')	('NumCatalogPurchases',)	0.5	0.37	*32*

با تحلیل قوانین متوجه میشویم که اکثر قوانین از هر دو طرف برقرارند اما با تحلیل مقدار confidence آنها، قانونی را انتخاب میکنیم که confidence بالاتری دارد.

مثلا قانون زیر از هر دو طرف برقرار است:

Education_Graduation <-> NumCatalogPurchases, NumWebPurchases

اما با بررسی confidence متوجه میشویم که قانون زیر مطمئن تر است:

Education_Graduation -> NumCatalogPurchases, NumWebPurchases

با بررسی این قوانین متوجه میشویم که آنهایی که مثلا فارغ التحصیل هستند، هر سه نوع خرید را انجام میدهند.

حال با ابر کلمات این ارتباط را بررسی میکنیم.

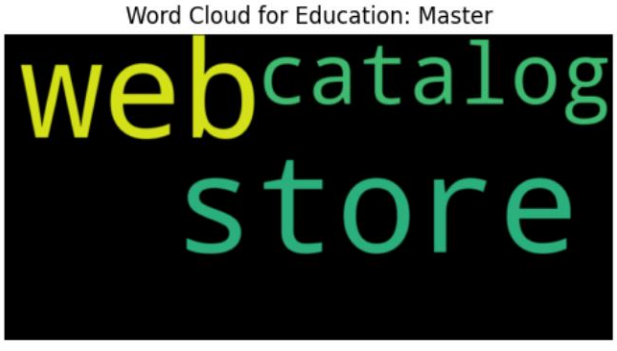
برای این کار ابتدا ستون های NumCatalogPurchases و NumWebPurchases و NumWebPurchases را به صورت باینری تبدیل میکنیم و در ستون های جدید آنها را ذخیره میکنیم. به این معنی که اگر خریدی داشتند مقدار ۱ و در غیر این صورت مقدار صفر داخل ستون مدنظر قرار گیرد. پس از این کار ما روی مقادیر مختلف ستون تحصیلات گام های زیر را انجام میدهیم:

۱. ابتدا داده هایی که مقدار مد نظر را برای ستون Education دارند (مثلا سطر هایی که تحصیلاتشان PhD هست را انتخاب میکنیم).
۲. سپس تعداد کاربران را برای هر یک از روش های خرید انجام شده از فروشگاه کرده بدست می آوریم.
۳. ابر کلمات را با توجه به تعداد تکرار هر روش خرید نمایش میدهیم.

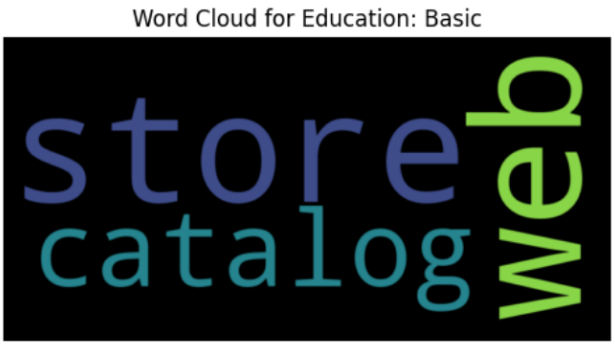
ابر کلمات برای هر یک از مقادیر تحصیلات به صورت زیر می باشد:

تصویر ۴- ابرکلمات مربوط به نحوه خرید افراد فارغ التحصیل از فروشگاه

تصویر ۵- ابرکلمات مربوط به نحوه خرید افراد PhD از فروشگاه



تصویر ۶- ابرکلمات مربوط به نحوه خرید افراد Master از فروشگاه



تصویر ۷- ابرکلمات مربوط به نحوه خرید افراد Basic از فروشگاه

تصویر ۸- ابرکلمات مربوط به نحوه خریده افراد 2n Cycle از فروشگاه

هر چه در ابر کلمات، یک کلمه بزرگتر باشد، تعداد تکرار آن بیشتر است در نتیجه از این طریق میتوانیم بفهمیم که در هر سطح تحصیل کدام روش خرید بیشتر رایج است.

مثلا برای افراد فارغ التحصیل خرید حضوری بیشتر از بقیه روش ها رایج است. برای بقیه سطوح تحصیل نیز به همین روش میتوانیم تحلیل را انجام دهیم.

۳-۲-۲- بخش دوم

برای این بخش ما میخواهیم الگو های پرتکرار برای ستون های Education و NumWebPurchases و NumCatalogPurchases و NumStorePurchases را پیدا کنیم تا بعد از آن در بخش اول تسک های نمره اضافه با استفاده از قوانین انجمنی ارتباط بین سطح تحصیلات و نحوه خرید از فروشگاه را بررسی کنیم.

برای این ستون ها لازم است ابتدا یکسری پیش پردازش روی داده ها انجام دهیم. ابتدا باید مقادیر ستون Education را که مقادیر ترتیبی هست را به صورت باینری تبدیل کنیم همچنین مقادیر ۳ ستون دیگر را به این صورت باینری میکنیم که اگر مقدار آنها صفر بود همان صفر و در غیر انی صورت مقدار ۱ باشد که به معنای انجام خرید به صورت اینترنتی یا با کاتالوگ و یا حضوری می باشد. حال با min_support = 0.3 الگوریتم apriori را روی این ۴ ستون اجرا میکنیم تا الگو های پرتکرار را پیدا کنیم که به صورت زیر می باشد:

جدول ۸- الگوهای پرتکرار با min_support=0.3

itemsets	support	#
(NumWebPurchases)	0.97	*1*
(NumCatalogPurchases)	0.73	*2*
(NumStorePurchases)	0.99	*3*
(Education_Graduation)	0.50	*4*
(NumCatalogPurchases, NumWebPurchases)	0.73	*5*
(NumStorePurchases, NumWebPurchases)	0.97	*6*
(Education_Graduation, NumWebPurchases)	0.49	*7*

(NumStorePurchases, NumCatalogPurchases)	0.73	*8*
(Education_Graduation, NumCatalogPurchases)	0.37	*9*
(NumStorePurchases, Education_Graduation)	0.49	*10*
(NumStorePurchases, NumCatalogPurchases, NumWebPurchases)	0.73	*11*
(Education_Graduation, NumCatalogPurchases, NumWebPurchases)	0.37	*12*
(NumStorePurchases, Education_Graduation, NumWebPurchases)	0.49	*13*
(NumStorePurchases, Education_Graduation, NumCatalogPurchases)	0.37	*14*
(NumStorePurchases, Education_Graduation, NumCatalogPurchases, NumWebPurchases)	0.37	*15*