

بسم الله الرحمن الرحيم



عنوان

## داده کاوی در پژوهش بازار

نگارش

علیرضا ایرانی

استاد راهنمای

دکتر محمدرضا فقیهی

دانشگاه شهید بهشتی

دانشکده علوم ریاضی

گروه آمار

## فهرست مطالب

6 .....	فصل اول : پژوهش بازار
6 .....	مقدمه
7 .....	برند چیست؟
7 .....	آگاهی از برند
8 .....	تفاوت بازاریابی پیشگویانه و بازاریابی سنتی
10 .....	پژوهش کمی و کیفی
10 .....	کارهایی که پژوهش بازار برای ما انجام می دهد
13 .....	آشنایی با برخی نمودارهای رایج پژوهش بازار
13 .....	هرم و قیف سلامت برند
14 .....	نمودار عنکبوتی
15 .....	نمودار میله ای انواع آگاهی
15 .....	نمودار میله ای انواع مصرف
16 .....	علاقه به برند
16 .....	ارزیابی قیمت
17 .....	کارایی برند
18 .....	مشتریان وفادار
20 .....	فصل دوم : داده کاوی
20 .....	مقدمه
21 .....	روش های رده بندی و پیشگویی

21	درخت رگرسیون
22	رگرسیون
24	K نزدیک ترین همسایه
25	شبکه عصبی
25	ماشین بردار پشتیبانی (SVM)
26	تحلیل تشخیصی (discriminant analysis)
27	سایر روش‌های داده کاوی
27	خوشه بندی
27	قواعد پیوند
29	فصل 3 : کاربرد داده کاوی در پژوهش بازار
29	مقدمه
35	تحلیل اکتشافی
35	الف) نمودارهای تک متغیره
52	ب) هیستوگرام انوع هزینه و درآمد ماهانه
56	ج) نمودارهای دو متغیره
68	د) نمودار پراکنش
73	ه) نمودارهای سه بعدی
78	درخت رده‌بندی
79	ماتریس درهم ریختگی داده های آموزشی
80	ماتریس درهم ریختگی برای داده‌های اعتبار سنجی
81	درخت هرس شده

83 .....	ماتریس درهم ریختگی داده‌های آموزشی درخت هرس شده
83 .....	ماتریس درهم ریختگی داده‌های اعتبارسنجی درخت هرس شده
84 .....	مدل رگرسیون ترتیبی
84 .....	ماتریس برهم ریختگی بر روی داده‌های آموزشی
85 .....	ماتریس برهم ریختگی بر روی داده‌های اعتبارسنجی
86 .....	روش SVM
87 .....	ماتریس برهم ریختگی بر روی داده‌های آموزشی
88 .....	ماتریس برهم ریختگی بر روی داده‌های اعتبارسنجی
90 .....	شبکه عصبی
93 .....	برازش مدل شبکه عصبی و ماتریس برهم ریختگی بر روی داده‌های آموزشی
94 .....	برازش مدل و ماتریس درهم ریختگی بر روی داده‌های اعتبارسنجی
95 .....	تحلیل تشخیصی
96 .....	برازش مدل تحلیل تشخیصی و ماتریس برهم ریختگی بر روی داده‌های آموزشی
98 .....	برازش مدل تحلیل تشخیصی و ماتریس برهم ریختگی بر روی داده‌های اعتبارسنجی
99 .....	روش k امین نزدیک ترین همسایه
101 .....	مدل بر روی داده‌های اعتبارسنجی
102 .....	مدل بر روی داده‌های آزمون
103 .....	تفسیر و بررسی نتایج و برازش بهترین مدل
104 .....	منابع
105 .....	پیوست
105 .....	دستورات بخش تصویری سازی

125 .....	دستورات بخش درخت رگرسیون
129 .....	دستورات بخش SVM
130 .....	دستورات بخش KNN
132 .....	دستورات بخش شبکه عصبی
135 .....	دستورات بخش رگرسیون ترتیبی
139 .....	دستورات بخش تحلیل تشخیصی

# فصل اول : پژوهش بازار

## مقدمه

قبل از هر چیزی به بررسی دو کلمه پژوهش و بازار بپردازیم. پژوهش یعنی جستجوی منظم اطلاعات جهت دستیابی به اهداف مشخص برای یافتن نتایج جدید است و بازار به مجموع مشتریان بلقوه و ب فعل یک محصول (کالا یا خدمات) گفته می‌شود.

یکی از اصلی‌ترین وظایف مدیریت، گرفتن تصمیم درست است و لازمه گرفتن یک تصمیم درست، داشتن داده‌های درست و دانش استفاده از این داده‌ها است. شرکت‌ها باید مشتریان و رقایشان را بشناسند و با روندها و سایر جنبه‌های بازار همگام شوند. به طور مثال ممکن است لازم باشد که دلیل فروش نرفتن محصولی را بدانند یا درباره ارائه خدمات یا محصول جدیدی تصمیم بگیرند و یا مقبولیت یک تبلیغ را پیش از نمایش، ارزیابی کنند. یکی دیگر از مزایای پژوهش بازار شناسایی موقعیت‌های خوب در زمان مناسب و ارائه محصولات متناسب با نیاز و سلیقه مشتریان سبب افزایش فروش شرکت‌ها و موسسات خواهد شد. مثلاً فردی قصد دارد رستورانی را تاسیس کند، نیاز دارد که اطلاعاتی راجع به میزان استقبال مشتریان از محصولات یا خدمات خود و یا بازار رقبا داشته باشد. پژوهش بازار این امکان را فراهم می‌سازد. انجام فرآیند پژوهش بازار به صورت دوره‌ای سبب می‌شود تا روندها و تاثیرات مثبت و منفی عوامل مختلف بر روی آگاهی، مصرف و خرید، شناسایی و در جهت آن اقدامات متناسب صورت گیرد.

در واقع همه این اقدامات با هدف کاهش ریسک صورت می‌گیرد و هر چه شناخت شما از بازاری که در آن مشغول هستید بیشتر باشد تصمیم‌ها با اطمینان بیشتر و ریسک کمتر گرفته می‌شود. همه این‌ها از جمله مزایای استفاده از پژوهش بازار هست که اکنون هر مدیری مستلزم استفاده از آن می‌باشد.

از طرف دیگر کیفیت اطلاعاتی که از طریق پژوهش بازار، جمع آوری می‌شود بسیار دقیق‌تر و مشرح‌تر از هرگونه اطلاعاتی است که به صورت پراکنده از منابعی که دارای اعتبار کمتری است بدست می‌آید. قبل از اینکه ادامه بدهیم نیاز است که با یکسری واژه‌های کاربردی درباره این موضوع بپردازیم

## برند چیست؟

قبل از آن که بگوییم برند چیست بهتر است بگوییم چه چیزهایی برند نیست خیلی‌ها برند را با محصول یا لوگو اشتباه می‌گیرند و یا بعضی از مشاوران تبلیغاتی می‌گویند برند تبلیغ و تاثیرگذاری بر مخاطب است اما هیچ یک از این‌ها درست نیست. برند یک نتیجه است، حس درونی مخاطب درباره محصول، سابقه از کسب و کار است. درواقع برندهای چگونگی تاثیر محصول که ارائه کردیم، طراحی و پیام محصولی که تدوین کردیم، فرهنگی که ایجاد کردیم، ظاهر و شیوه رفتار کارمندان است. برند تجربه و ادراکی است که از سازمان، محصول یا شخص در ذهن ما شکل گرفته است به عبارت دیگر ادراکی از موجودی پویا است در لایه‌های پنهان ذهنی است.

## آگاهی از برند

آگاهی از برند<sup>1</sup> یکی از اصلی ترین شاخص‌های کسب و کار است که بر اساس آن مشخص می‌شود چند نفر از جامعه هدف برند شما را می‌شناسند. در اکثر موقعیت‌های مشتری در لحظه آشنایی با برند، از آن خرید نمی‌کند. ترجیح مشتری بر خرید از برند آشناست به این دلیل است که مدیران شرکت‌ها سعی در افزایش آگاهی برند خود هستند.

آگاهی از برند شامل : اولین برند که به ذهن می‌رسد<sup>2</sup> که در این نوع از آگاهی پرسشگر بدون اینکه نام برندی را ذکر کند از پاسخگو می‌خواهد تا اولین برندی که به ذهنش می‌رسد را بگوید. این نوع سوال به صورت تک پاسخی است و پاسخگو فقط مجاز است نام یک برند را بگوید و در صورتی که نام چند برند مطرح شد پرسشگر می‌بایست اولین برند را یادداشت کند. نوع دیگر آگاهی، آگاهی بدون کمک<sup>3</sup> است که پرسشگر باز بدون اینکه نام برندی را ذکر کند از پاسخگو می‌خواهد تا هر تعداد برندی را که می‌شناسد ذکر کند این نوع سوال چند پاسخی می‌باشد و به همراه اولین برند که به ذهن می‌رسد به عنوان آگاهی بدون کمک شناخته می‌شود و نوع اخر آگاهی،

Brand Awareness<sup>1</sup>

Top of mind<sup>2</sup>

Unaided Awareness<sup>3</sup>

آگاهی کل<sup>۴</sup> است. در این سوال پرسشگر نام برندها را برای پاسخگو می‌خواند و از او می‌خواهد برندهایی که اسمشان را شنیده است را بگوید حتی اگر از ان محصول استفاده نکرده باشد.

## تفاوت بازاریابی پیشگویانه<sup>۵</sup> و بازاریابی سنتی

اساس بازاریابی سنتی<sup>۶</sup> ارزیابی یک بازار برای یک محصول یا خدمات متمرکز است و این نوع بازاریابی برنده و محصول مبنای است در واقع هدف فهمیدن میزان آگاهی از برنده و میزان استفاده و ... است اما بازاریابی پیشگویانه، مبنا مشتری است، این روش بازاریابی شامل استفاده از تجزیه و تحلیل داده‌ها است برای تعیین اینکه کدام استراتژی‌ها و اقدامات بازاریابی بیشترین احتمال موفقیت را دارند. در این روش تلاش بر این است که رفتار مشتری پیشگویی شود و از قواعد پیوند که در فصل بعدی درباره آن بیشتر توضیح داده می‌شود نیز استفاده می‌شود. از عوامل مهم در بازاریابی پیشگو این است که مشتریان خواهان یک رویکرد شخصی و یکپارچه‌تر هستند آنها از طریق بسیاری از کانال‌ها با بازاریابی و فروش تعامل دارند و همچنین در این روش می‌توان با استفاده از داده‌های موجود مشتریان و با استفاده از روش‌های نمونه‌گیری، الگوهای را بشناسیم و از داده‌های مشتری در تقاطع دنیای فیزیکی و دیجیتال استفاده کنیم.

### مزایای بازاریابی پیشگو

۱. پیشبینی رفتار مشتری و احتمال خرید آینده مشتریان بلقوه
۲. به جای پیدا کردن مشتری برای محصول، محصولی طراحی می‌شود که مشتری می‌خواهد در آینده بخرد
۳. به جای ماسکسیم کردن فروش، تمرکز شرکت بر بهینه سازی ارزش عمر مشتری
۴. به جای سازمان‌دهی خط تولید و کانال‌ها، شرکتها تلاش بر سازمان‌دهی مشتری دارند
۵. ارتباطات با مشتریان بسیار هدفمندتر می‌شوند.
۶. افزایش وفاداری مشتریان از طریق شخصی سازی کردن تجارت که به مشتری انتقال داده می‌شود
۷. افزایش دقت در شناسایی مشتریان هدف

---

Total Awareness<sup>4</sup>  
Predictive Marketing<sup>5</sup>

آرایشگر خود را درنظر بگیرید. او اطلاعات زیادی از شما دارد. او می‌داند که به چه مدل موبی علاقه دارد و یا اطلاعاتی درباره شعل و خانواده شما دارد. این اطلاعات ارتباط شما با آرایشگرتان را تحقیم می‌بخشد. شما می‌نشینید و او کارش را می‌کند و شما یک گفتگوی دلپذیر را تجربه می‌کنید. آرایشگرتان تعداد مشتریان کمی دارد اما تحلیل و پردازش داده‌های شرکت بزرگ با مشتریان میلیونی بدون نرم افزارها غیر ممکن است.



## پژوهش کمی و کیفی<sup>۷</sup>

پژوهش کیفی بر اساس ترجیحات، احساسات و نظرات مشتریان است و هدف ان روش نمودن پیچیدگی‌های یک مسئله است و نیازمند تعامل مستقیم با پاسخ‌دهندگان است که معمولاً با انجام مصاحبه‌های عمیق که بیشتر سوالات، مربوط به سوالات باز می‌باشد صورت می‌گیرد. اما پژوهش کمی بر اساس اندازه‌گیری‌های دقیق جنبه‌های بازاریابی مانند حجم بازار، سطح توزیع، روندهای فروش است با هدف پیش‌بینی، تشریح و توصیف پدیده‌ها که قابل تجزیه و تحلیل باشد. معمولاً به صورت سوالات بسته از طریق مصاحبه‌های حضوری یا آنلاین تکمیل می‌شود. معمولاً تعداد نمونه‌ها در پژوهش کیفی کمتر از پژوهش کمی است اما هزینه هر واحد نمونه در پژوهش کیفی به مراتب بیشتر از پژوهش کمی است.

### کارهایی که پژوهش بازار برای ما انجام می‌دهد

عمده مطالعات که در زمینه پژوهش بازار صورت می‌گیرد با هدف شناخت مشتریان از نظر سبک زندگی، عادات خرید و مصرف، نگرش به ابعاد مختلف محصولات یا خدمات می‌باشد. سنجش میزان رضایت مشتری یکی دیگر از اهداف اصلی پژوهش بازار است و شامل ابعاد مختلف محصول یا خدمات، از قیمت و کیفیت تا سهولت دسترسی و یا اندازه محصول می‌باشد. مدیران معمولاً علاقمند هستند اطلاعات یک پژوهش بازار به تفکیک شهر، جنسیت، گروه‌بندی سنی، طبقه اقتصادی اجتماعی و داشتن یا نداشتن مسئولیت خرید ملاحظه کنند، تا با این کار به مقایسه محصول خود در بخش‌های مختلف پیردازند به طور مثال در یک پژوهش پژوهش بازار شرکت تولید آبمیوه متوجه می‌شود که میزان مصرف آبمیوه در یک شهر با اختلاف معناداری کمتر از سایر شهرهای است و از سوی دیگر متوجه می‌شود گروه سنی 18 تا 25 سال با اختلاف معناداری مصرف بیشتری از آبمیوه تولیدی این کارخانه را نسبت به گروه‌های سنی دیگر دارد و از انجایی که می‌داند میانگین سنی این شهر پایین تر از شهرهای دیگر است و مدیر تصمیم می‌گیرد هزینه تبلیغ بیشتری را برای این شهر بکند.

حال یک شرکت آبمیوه الف برای یک پژوهش پژوهش بازار نیاز دارد چه سوالاتی طراحی کند به طور مثال:

1. هر چند وقت یکبار آبمیوه را خریداری می کنید؟
2. در هر بار خرید معمولاً چه سایزی از آبمیوه خریداری می کنید؟
3. معمولاً آبمیوه را از کجا خریداری می کنید؟
4. در چه زمان هایی آبمیوه مصرف می کنید؟ قبل و بعد باشگاه، در دانشگاه و یا در مهمانی؟
5. چه برندهای آبمیوه را می شناسید؟
6. کدام برنده آبمیوه را اغلب اوقات مصرف می کنید؟
7. قبل از برنده اغلب اوقات چه برنده مصرف می کردید؟ دلیل این تغییر چه بوده است؟
8. مهم ترین عامل انتخاب آبمیوه از نظر شما چیست؟ دیگر چه عواملی؟
9. بگویید بهترین مکان برای تبلیغات محصولات انواع آبمیوه از نظر شما کدام رسانه است؟

همه اینها سوالات مهمی است که با تحلیل شان می تواند مبنای تصمیم گیری های شرکت ها قرار گیرد و مشکلات بسیاری را حل کند.

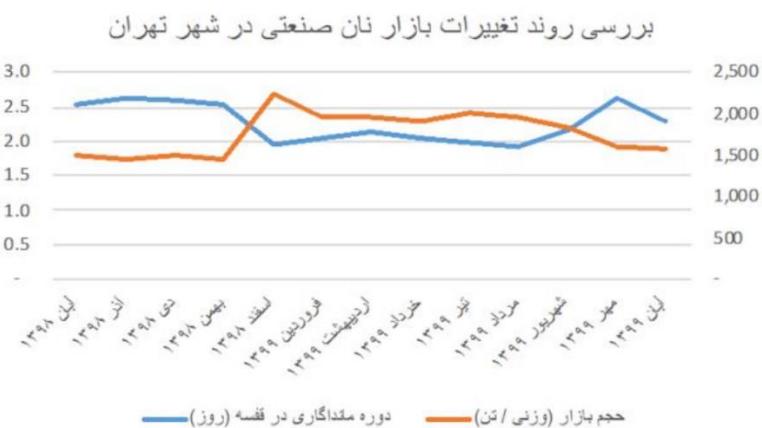
فرض کنید یک شرکت تولید کننده شامپو هزینه زیادی را صرف تولید بسته بندی جدیدش کرده و در این پروژه متوجه می شود که مهمترین عوامل انتخاب شامپو از نظر پاسخگویان به ترتیب قیمت، نرم کنندگی، رایحه، خاصیت ضدشوره و درنهایت بسته بندی است. یعنی از نظر مشتریان بسته بندی محصول در اولویت اخر قرار می گیرد و این شرکت می تواند هزینه بسته بندی را کاهش دهد و از قیمت بکاهد تا میزان فروش محصولش بیشتر شود و یا در روش های تبلیغ متوجه شود که بیلبورد کمترین میزان اثرگذاری را دارد و شرکت می تواند میزان هزینه تبلیغات این بخش را کاهش دهنده. از مثال های بالا کاملاً متوجه شدیم که یکی از کاربردهای مهم پژوهش بازار بهینه سازی هزینه است که از دیدگاه مدیران اهمیت ویژه ای دارد.

از دیگر مزایای پژوهش بازار کشف فرصت ها است. برای مثال یک شرکت تولید کننده لوازم بهداشتی می خواهد بداند که آیا میزان مصرف لوازم بهداشتی قبل و بعد از شیوع بیماری کرونا تغییری داشته و اگر تغییری داشته این تغییر متعلق به کدام محصولاتش بوده است. به کمک اطلاعاتی که پژوهش بازار در اختیار این شرکت قرار می دهد مدیر شرکت می تواند درباره میزان تولید محصولاتش در آینده و یا تولید محصول جدید مطابق با نیاز بازار اقدام کند.

بدون شک شیوع ویروس کرونا تاثیر بسیاری بر بازارهای جهانی و داخلی داشته است. نان صنعتی یکی از بخش های بازار است که پس از اعلام شیوع ویروس کرونا در کشور دچار نوسانات جدی شده است. طبق پژوهش صورت گرفته پیش از نقطه اوج شیوع بیماری کرونا (اسفند ۱۳۹۸) حجم فروش نان صنعتی با روندی ثابت ماهیانه در حدود ۱۵۰۰ تن و دوره ماندگاری در قفسه ۲.۵ روز بوده است. اما بازار نان صنعتی به محض اعلام ورود ویروس به کشور و افزایش نگرانی ها و اعلام رعایت پروتکل های بهداشتی دچار نوسان شده است.

پژوهش نشان می‌دهد نتایج دوره اسفند ۱۳۹۸ شاهد رشد ۲۳ درصدی در حجم فروش این محصول و کاهش دوره ماندگاری در قفسه به کمتر از ۲ روز بوده است.

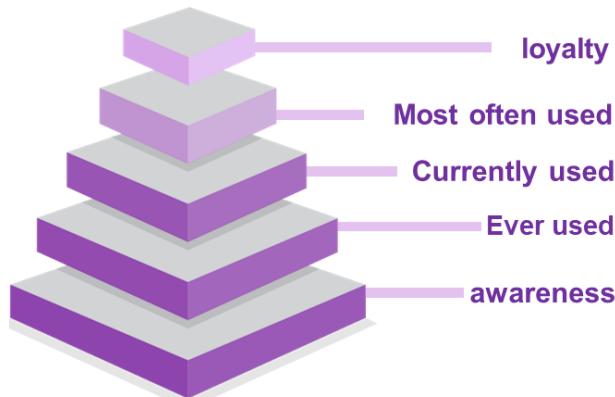
پس از اینکه پیک اول شیوع ویروس سپری شد و آمار مرگ و میر ناشی از کرونا کاهش پیدا کرد. مصرف کنندگان به سرعت تغییر رفتار داشته و روند خریدشان در بازار نان صنعتی مسیر نزولی و بازگشت به شرایط قبلی را تجربه می‌کند. پژوهش نشان داد: در ماههای بعد از پیک اول یعنی به مدت ۶ ماه (تا انتهای مرداد ۱۳۹۹) با افت ملایم حجم فروش و افزایش دوره ماندگاری محصول در قفسه‌ها بیش از ۲ روز روبرو هستیم. بنابراین در شهریور و مهر ۱۳۹۹ بازار در مسیر بازگشت به شرایط اولیه خود بوده تا اینکه دوره آمار مرگ و میر افزایش پیدا کرده و کشور وارد پیک دوم شیوع ویروس کرونا می‌شود. در این دوره نیز بازار نان صنعتی نوسانات جدیدی را تجربه می‌کند. پژوهش نشان می‌دهد، در آبان ماه ۱۳۹۹ با افزایش نگرانی‌های همه گیری بیماری کووید - ۱۹ روند کاهشی حجم فروش متوقف شده و دوره ماندگاری با اندکی تغییر به کمی بیش از ۲ روز رسیده است. چنین نوساناتی به وضع در نمودار گزارش پژوهشی بررسی دوره‌ای بازار خرده فروشان نان صنعتی شهر تهران که اطلاعات آن از ۲۴ شهر ایران جمع آوری شده به وضع دیده می‌شود. این اطلاعات برای برنامه‌ریزی و تدوین استراتژی فعالان حوزه بازار نان صنعتی مفید خواهد بود.



## آشنایی با برخی نمودارهای رایج پژوهش بازار

### هرم و قیف سلامت برنده

نمودار زیر یکی از مهم ترین نمودارهای مورد نیاز شرکت ها است که ان را هرم سلامت برنده می نامند. نمودار زیر مربوط به یک شرکت تولیدی لوازم خانگی است که به ترتیب از پایین به بالا میزان آگاهی کل، مصرف تابحال، مصرف اخیر، مصرف اغلب اوقات و وفاداری مشتری به برنده را نمایش می دهد. هر چه این نمودار به حالت مستطیل تر باشد بهتر است یعنی آگاهی از برنده منجر به مصرف بیشتر و وفاداری بیشتر (توصیه برنده به دیگران) می شود



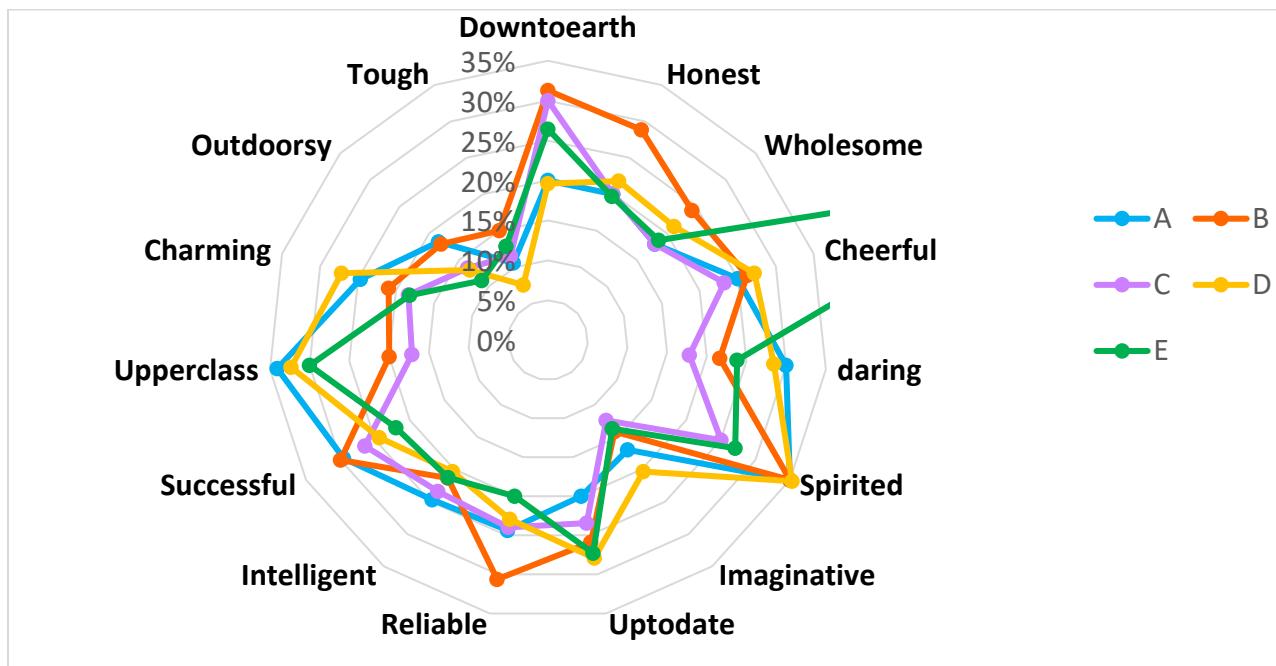
Brand pyramid



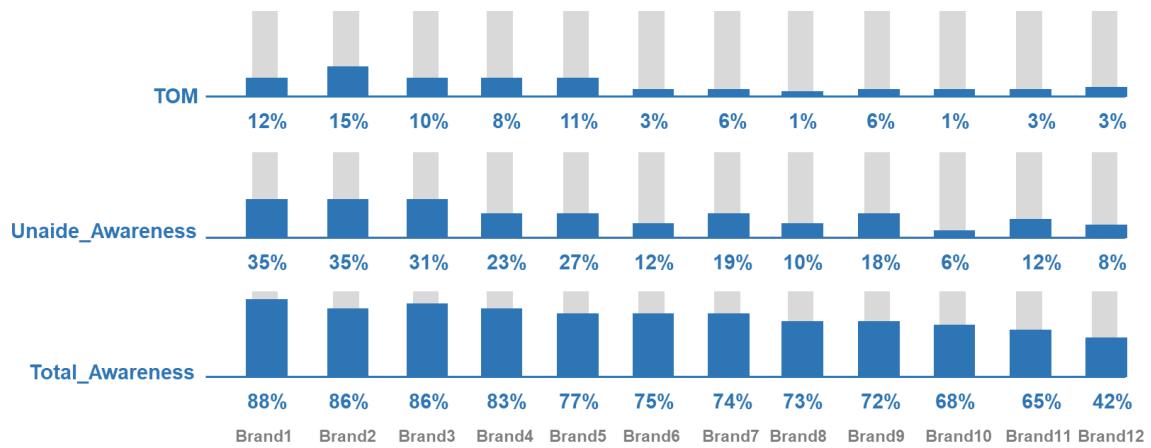
Brand funnel

## نمودار عنکبوتی

به کمک نمودار عنکبوتی می‌توان صفت‌های مختلف را برای برندهای اصلی بازار بررسی کرد.



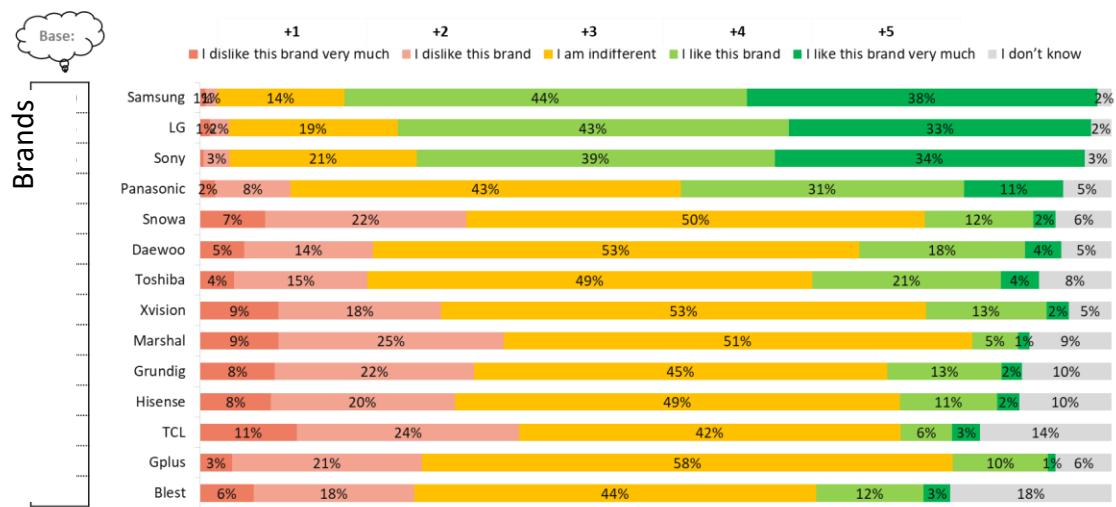
نمودار میله ای انواع آگاهی



نمودار میله ای انواع مصرف



## علاقه به برنده

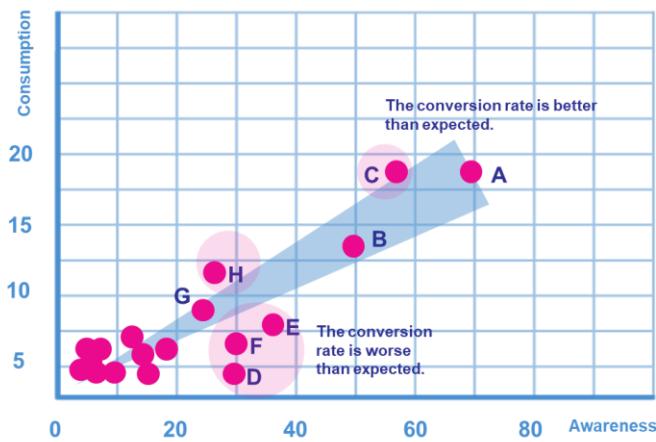


## ارزیابی قیمت

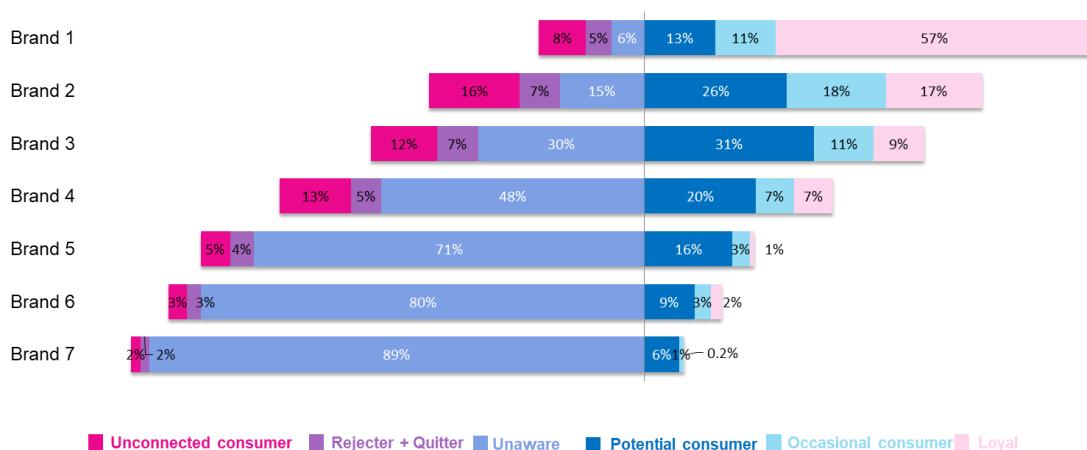


## کارایی برنده<sup>۸</sup>

نمودار آگاهی در مقابل مصرف است که به نمودار کارایی برنده شهرت دارد.



نمودار Brand Engagement<sup>9</sup>



Brand Efficiency<sup>8</sup>

## مشتریان وفادار

مهم ترین اصل در فروش کالا جلب اعتماد خریدار است. یکی از موثرترین روش های جذب مشتری این است که برنده شما از طریق انتقال فردی که تجربه استفاده محصول شما را داشته به دیگران است. در پژوهش بازار این افراد ترویج دهنده را مشتریان وفادار<sup>۹</sup> می نامند. برخی وفاداری را همان ارتباط عاطفی بین مشتری و برنده و کسب و کار دانسته اند، تعهد مشتری به خرید همیشگی از یک برنده و در جای دیگر مشتریانی که با معرفی محصول به دیگران بیشترین سود را وارد چرخه کسب و کار می کنند.

چگونه وفاداری مشتری را اندازه گیری کنیم؟

برای محاسبه این شاخص از پاسخگویان می پرسیم که چقدر احتمال دارد که برندهای زیر را به دیگران توصیه کنید. پاسخگویان بر اساس پاسخی که می دهند به سه گروه : ترویج دهنده<sup>۱۰</sup> (۱۰-۹-۸) و منفعل<sup>۱۱</sup> (۷-۶-۵) و دفع کننده<sup>۱۲</sup> (۴-۳-۲) تقسیم بندی می شوند مانند

جدول زیر:

Base: 10!	Brands	I won't recommend at all	1 2 3 4 5 6						7 8		I will definitely recommend	I don't know	
			1	2	3	4	5	6	7	8			
10!	Sung	1%	1%	1%	2%	3%	5%	6%	12%	17%	20%	32%	0%
10!	sonic	2%	1%	1%	2%	3%	6%	8%	10%	21%	17%	27%	0%
10!	va	2%	1%	2%	2%	4%	6%	8%	12%	16%	17%	30%	1%
96	voo	5%	5%	5%	7%	9%	14%	10%	13%	13%	8%	8%	2%
85	oo	18%	11%	12%	10%	10%	10%	8%	8%	5%	2%	3%	3%
74	iba	9%	9%	11%	12%	11%	12%	9%	8%	7%	4%	5%	4%
64	on	13%	11%	8%	11%	9%	13%	9%	8%	8%	3%	3%	4%
60	hal	18%	10%	9%	12%	11%	13%	7%	9%	3%	2%	2%	3%
53	dig	22%	14%	13%	10%	9%	11%	5%	5%	2%	1%	2%	5%
18	nse	20%	13%	11%	10%	8%	11%	5%	4%	2%	4%	3%	7%
16	s	17%	11%	10%	17%	7%	10%	6%	10%	2%	2%	4%	4%
16	se	25%	12%	9%	9%	10%	11%	5%	8%	3%	1%	2%	4%
11	o	12%	8%	15%	9%	11%	19%	9%	5%	3%	3%	2%	5%
34	o	18%	12%	3%	15%	6%	18%	12%	3%	6%	0%	6%	3%

Customer Loyalty<sup>9</sup>  
Promoters<sup>10</sup>  
Detractors<sup>11</sup>

## محاسبه شاخص خالص ترویج دهنده‌گان<sup>12</sup>

$$NPS = \text{Promoters\%} - \text{Detractors\%}$$

با محاسبه این شاخص برای هر برنده شاخص خالص ترویج دهنده‌گان هر برنده مشخص می‌شود البته نیاز است که شاخص محاسبه شده برای هر برنده با یک واحد تحت عنوان شاخص ترویج دهنده‌گی بازار مقایسه شود. بدین منظور یکسان در نظر گرفتن تمام برنده‌ها با استفاده از فرمول فوق میزان خالص ترویج دهنده‌گان کل مشخص می‌شود. و برنده‌های که NPS آنها بیشتر از NPS کل باشد برندهای ترویج شونده محسوب می‌شوند.

## فصل دوم : داده کاوی

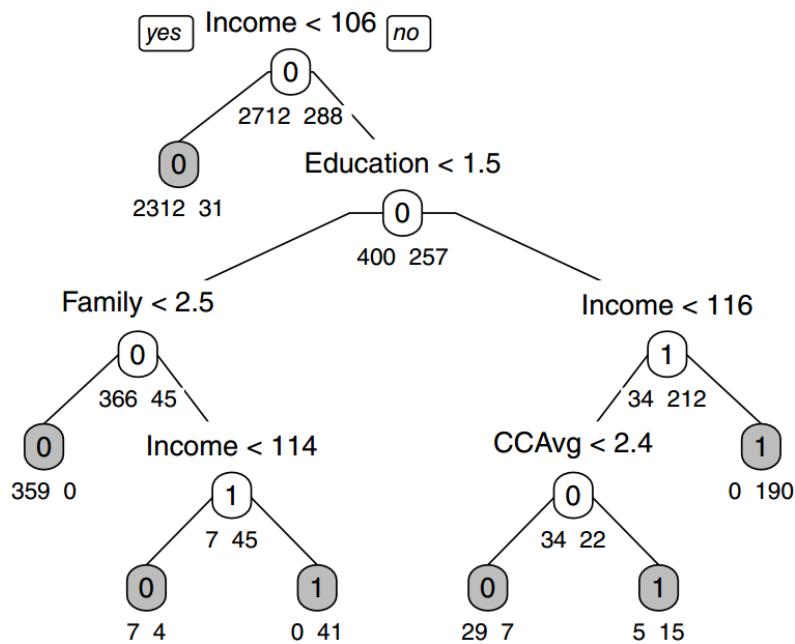
### مقدمه

داده کاوی شامل مجموعه‌ای از روش‌های تحلیل خودکار کسب و کار است که از شمارش، فنون توصیفی، گزارش‌دهی و روش‌های مبتنی بر قواعد کسب و کار فراتر می‌رود به طور کلی هدف داده کاوی رده بندی و پیشگویی است و به شدت در دهه اخیر در حوزه‌های مختلف به کار گرفته شده است. بی‌درنگ تصمیمات مربوط به حوزه پژوهش بازار با حجم اطلاعات زیاد نیاز به تکنیک‌های پیشرفته‌ای از جمله داده کاوی دارد. در این کتاب تلاش بر این است که به بررسی مفهوم داده کاوی و نقش آن در پژوهش بازار پرداخته شود. داده کاوی از تلفیق آمار و یادگیری ماشین شکل گرفته است. آمار یک متوسط و یک نتیجه می‌دهد، به طور مثال به ازای افزایش یک واحد قیمت، دو واحد تقاضا کاهش می‌یابد. ولی در یادگیری ماشین در مورد تک افراد بحث می‌کند یعنی برای هر خریدار نتیجه متفاوتی متصور می‌شود یادگیری ماشین، الگوریتم‌هایی است که مستقیماً از داده‌ها، بصورت لایه‌بندی یا تکراری یاد می‌گیرند. به طور کلی داده کاوی به دو بخش راهنماییده و ناراهنماییده تقسیم بندی می‌شود یادگیری راهنماییده به فرایند آماده کردن یک آلگوریتم اتلاق می‌شود که با ثابت‌هایی که در آنها یک متغیر خروجی مورد نظر معلوم بوده و آلگوریتم «یاد می‌گیرد» چگونه این مقدار را با داده‌های جدید وقت خروجی معلوم نیست، پیشگویی کند اما یادگیری ناراهنماییده به تحلیلی اتلاق می‌شود که در آن کوشش برای یادگیری برخی چیزها درباره‌ی داده‌ها، به غیر از پیشگویی یک مقدار خروجی مورد نظر (مثالاً آیا در خوشه می‌افتد)، صورت می‌گیرد. به طور کلی داده‌ها را به سه بخش داده‌های آموزشی، داده‌های اعتبارسنجی، داده‌های آزمون تقسیم بندی می‌کنند. داده‌های آموزشی بخشی از داده‌ها است که برآش مدل روی آن صورت می‌گیرد و داده‌های اعتبارسنجی بخشی از داده‌ها که برای ارزیابی میزان خوب بودن مدل برآش شده، اصلاح برخی مدل‌ها و انتخاب بهترین مدل از میان مدل‌ها استفاده می‌شوند و در نهایت داده‌های آزمون بخشی از داده‌ها که فقط در پایان فرایند انتخاب و ساخت مدل، به منظور ارزیابی خوب بودن مدل نهایی مورد استفاده قرار می‌گیرند.

## روش های رده بندی و پیشگویی

### درخت رگرسیون

یک تکنیک رده بندی است که در طیف وسیعی از موقعیت ها خوب عمل کرده و نیازی به تلاش چندان تحلیل گر ندارد، در حالی که به راحتی توسط مصرف کننده تحلیلی قابل درک است. در رده بندی، متغیر برآمد یک متغیر رسته ای خواهد بود. یکی از دلایل محبوبیت درختان رده بندی این است که قوانین رده بندی قابل فهم را ارائه می دهند. ایده اصلی درخت تصمیم این است که افرادی را که شبیه به هم هستند را در یک رده قرار دهد. به عنوان مثال یک بانک برای اینکه مشخص کند به چه کسانی شرایط گرفتن وام را دارند از مدل درخت تصمیم استفاده کرده و بر اساس متغیرهای درآمد، تعداد اعضای خانواده، سطح آموزش و... مدل پیاده سازی شده است.



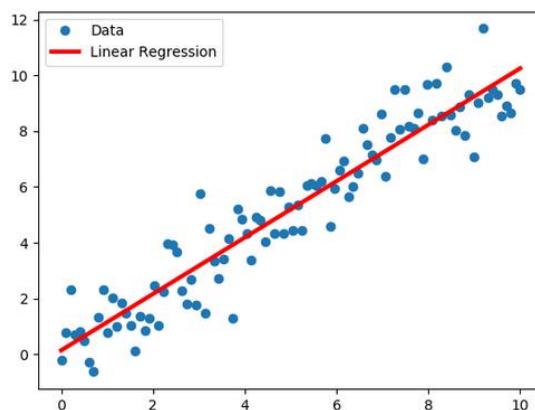
## رگرسیون

کلمه رگرسیون از لحاظ لغوی به معنای بازگشت است. تحلیل رگرسیونی یک فرآیند برآورد رابطه بین متغیرهای وابسته(پاسخ) و متغیرهای مستقل(پیشگو) است. در رگرسیون خطی بازش مدل خطی به مجموعه داده‌ها صورت می‌گیرد. به عبارت دیگر رگرسیون یک امید ریاضی شرطی است ( $E(Y|X)$ ) یعنی مقدار مورد انتظار  $Y$  به شرط اینکه مقدار  $X$  را بدانیم. از سوی دیگر می‌توان گفت مقادیر

$$\hat{Y} = HY \quad \text{پیش‌بینی شده } \hat{Y} \text{ تبدیل خطی از } Y \text{ است}$$

$$H = X(X^T X)^{-1} X^T$$

$$\hat{Y}_i = \alpha + \hat{\beta}_1 x_i + \cdots + \hat{\beta}_k x_k + \varepsilon_i \quad \text{نمایش مدل خطی}$$

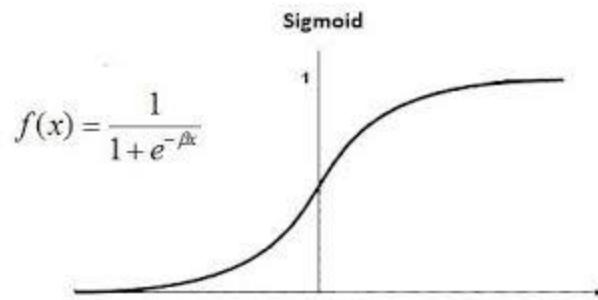


می‌خواهیم ضرایب بتا را طوری برآورد کنیم که ضریب همبستگی بین  $Y$  و  $\hat{Y}$  ماقزیم شود که توان دوم این ضریب همبستگی را ضریب تبیین ( $R^2$ ) می‌گوییم که عددی بین صفر و یک است. بدین منظور از روش OLS که همان مینیم کردن خطاهای استفاده است کنیم از انجایی که مقدار  $\epsilon_i$  می‌تواند کثبت یا منفی باشد، از تابعی از  $\epsilon_i$  استفاده می‌کنیم که همواره مثبت باشد قدر مطلق از آنجایی که مشتق پذیر نیست از توان دوم خطاهای استفاده می‌کنیم و ضرایب بتا را برآورد می‌کنیم

$$\begin{aligned} \hat{\alpha} &= \bar{y} - (\hat{\beta} \bar{x}), \\ \hat{\beta} &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

## رگرسیون لجستیک

در رگرسیون لجستیک متغیر پاسخ فقط دو مقدار ۰ و ۱ را اختیار می‌کند و به منظور رده‌بندی از آن استفاده می‌شود که در آن از تابع لجیت استفاده می‌کنیم



و به طور کلی

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_q x_q)}}$$

که بخت موفقیت به صورت زیر است

$$Odds(Y = 1) = \frac{p}{1 + p}$$

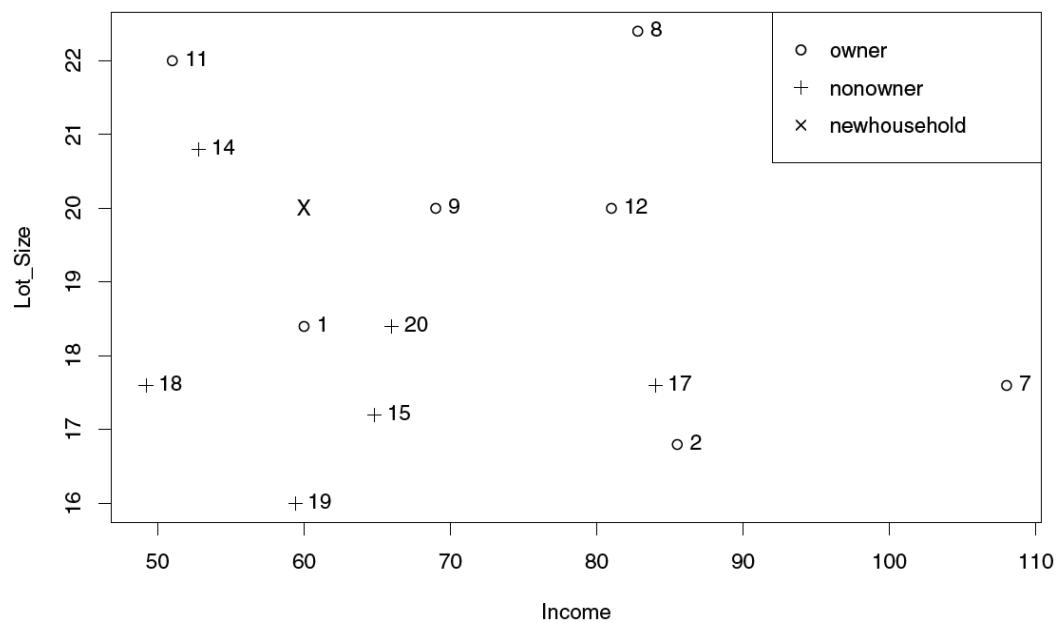
و در نتیجه مدل لجیت به صورت زیر می‌باشد

$$\log(Odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_q x_q$$

می‌توان این تفسیر را نیز داشت که با فزایش ۱ واحد مقدار  $x_1$  بخت موفقیت به اندازه  $e^{\beta_1}$  افزایش می‌یابد.

## K نزدیک ترین همسایه

ایده‌ی روش‌های k-نزدیک‌ترین همسایه عبارت است از شناسایی k ثبت در مجموعه داده‌های آموزشی که شبیه یک ثبت جدید که می‌خواهیم آن را رده‌بندی کنیم، هستند. سپس از این ثبت‌های مشابه، برای رده‌بندی ثبت جدید در یک رده، استفاده می‌کنیم. انتخاب دقیق و بهینه با اولین بررسی می‌تواند انجام شود و مقدار k بزرگ‌تر نویز را کاهش می‌دهد برای مثال یک شرکت ماشین چمن زنی k قصد دارد براساس درآمد و مساحت حیاط هر فرد پیش‌بینی کند که فرد جدید در رده خریدار قرار می‌گیرد یا خیر



در این روش در نقاط پیوسته از فاصله اقلیدسی و در نقاط گسسته از فاصله همینگ استفاده می‌کنیم

## شبکه عصبی

شبکه‌های عصبی مصنوعی یا شبکه‌های عصبی صناعی<sup>۱۳</sup> یا به زبان ساده‌تر شبکه‌های عصبی سیستم‌ها و روش‌های محاسباتی نوین برای یادگیری ماشینی، نمایش دانش و در انتها اعمال دانش به دست آمده در جهت بیش‌بینی پاسخ‌های خروجی از سامانه‌های پیچیده هستند. ایده اصلی این گونه شبکه‌ها تا حدودی الهام‌گرفته از شیوه کارکرد سیستم عصبی زیستی برای پردازش داده‌ها و اطلاعات به منظور یادگیری و ایجاد دانش می‌باشد. عنصر کلیدی این ایده، ایجاد ساختارهایی جدید برای سامانه پردازش اطلاعات است.

یک شبکه عصبی مصنوعی، از سه لایه ورودی، خروجی و پردازش تشکیل می‌شود. هر لایه شامل گروهی از سلول‌های عصبی (نورون) است که عموماً با کلیه نورون‌های لایه‌های دیگر در ارتباط هستند، مگر این که کاربر ارتباط بین نورون‌ها را محدود کند؛ ولی نورون‌های هر لایه با سایر نورون‌های همان لایه، ارتباطی ندارند. نورون کوچک‌ترین واحد پردازشگر اطلاعات است که اساس عملکرد شبکه‌های عصبی را تشکیل می‌دهد. یک شبکه عصبی مجموعه‌ای از نورون‌های است که با قرار گرفتن در لایه‌های مختلف، معماری خاصی را بر مبنای ارتباطات بین نورون‌ها در لایه‌های مختلف تشکیل می‌دهند. نورون می‌تواند یک تابع ریاضی غیرخطی باشد، در نتیجه یک شبکه عصبی که از اجتماع این نورون‌ها تشکیل می‌شود، نیز می‌تواند یک سامانه کاملاً پیچیده و غیرخطی باشد. در شبکه عصبی هر نورون به طور مستقل عمل می‌کند و رفتار کلی شبکه، برآیند رفتار نورون‌های متعدد است. به عبارت دیگر، نورون‌ها در یک روند همکاری، یکدیگر را تصحیح می‌کنند.

## ماشین بردار پشتیبانی (SVM)

این روش از جمله روش‌های نسبتاً جدیدی است که در سال‌های اخیر کارایی خوبی نسبت به روش‌های قدیمی‌تر برای طبقه‌بندی نشان داده است. مبنای کاری دسته‌بندی کننده SVM دسته‌بندی خطی داده‌ها است و در تقسیم خطی داده‌ها سعی می‌کنیم خطی را انتخاب کنیم که حاشیه اطمینان بیشتری داشته باشد. حل معادله پیدا کردن خط بهینه برای داده‌ها به وسیله روش‌های QP که روش‌های شناخته شده‌ای در حل مسائل محدودیت‌دار هستند صورت می‌گیرد. قبل از تقسیم خطی برای اینکه ماشین بتواند داده‌های با پیچیدگی بالا را دسته‌بندی کند داده‌ها را به وسیله تابع  $\phi$  به فضای با ابعاد خیلی بالاتر می‌بریم

## تحلیل تشخیصی (discriminant analysis)

تحلیل تشخیصی برای طبقه بندی پاسخگویان بر اساس مقادیر(کدهای) یک متغیر وابسته اسمی دو یا چند وجهی به کار می رود. در واقع در مواردی که متغیر وابسته اسمی و متغیر مستقل کمی باشد به منظور پیش بینی تغییرات متغیر وابسته از روی متغیرهای مستقل از تحلیل تشخیصی استفاده می شود. تحلیل تشخیصی روشی است که متغیرهای مستقل را برای ایجاد یک متغیر جدید ترکیب می کند که هر یک از پاسخگویان برای آن مقداری به دست می آورند. این متغیر جدید که تابع تشخیصی نامیده می شود به گونه ای محاسبه می شد که پاسخگویان را بر حسب مقداری که به دست می آورند در طبقات مختلف مختلف متغیر وابسته تفکیک کند. بنابراین تحلیل تشخیصی در صدد است تا ترکیبهای خطی بین متغیرهای مستقل را که قادرند به بهترین نحو گروههای پاسخگویان را از هم جدا کنند شناسایی کند. این ترکیب های خطی توابع تشخیصی نام دارند.

بطور مثال فرض کنید تولیدکننده ای در صدد است بداند آیا محصول جدید تولید شده مانند جاروبرقی نسبت به محصول قبلی از طرف مصرف کنندگان مورد استقبال قرار می گیرد یا خیر. برای این کار فروشنده باید افرادی را که حاضر به خرید هستند و افرادی که تمایلی به خرید ندارند را شناسایی کند. بدین منظور سه معیار دوام ، نحوه کار و شکل ظاهری را در سطح فاصله ای در نظر می گیرد. این متغیرها در واقع ملاک هایی هستند که خریدار بر مبنای آن محصول جدید را ارزیابی میکند. پاسخگو می تواند امتیازی بین ۰ تا ۱۰ را به هریک از معیارهای فوق اختصاص دهد. ترکیب وزنی این معیارها نشان خواهد داد که تا چه حد احتمال خرید کالای مورد نظر وجود دارد ، کدام معیار بیشتر در تصمیم به خرید مؤثر است و کدام عامل بیشتر خریداران را از کسانی که مایل به خرید نیستند تفکیک می کند. همبستگی بالای پاسخ متغیر وابسته (خرید یا عدم خرید) با هریک از معیارهای مورد نظر(دوام، نحوه کار و شکل ظاهری) نشان دهنده آن است که آن معیار برای تفکیک خریداران و غیر خریداران مناسب تر می باشد. به طور کلی می توان گفت تکنیک تحلیل تشخیصی چندگانه ابعاد و یا ویژگیهایی که در آن بیشترین تشخیص وجود دارد را در بین گروه ها شناسایی کرده و ضریب وزنی تشخیصی را برای هر متغیر به منظور انعکاس این تفاو ها بدست می دهد.

## سایر روش‌های داده‌کاوی

### خوشه‌بندی

خوشه‌بندی یکی از الگوریتم‌های بدون ناظر داده کاوی است. یکی از پرکاربردترین روش‌های خوشه‌بندی روش **k میانگین** است که براساس معیار شباهت بین نمونه‌ها به طور مثال تراکنش هر مشتری که با فاصله اقلیدسی محاسبه می‌شود، نمونه‌های یکسان در خوشه‌های یکسان قرار گرفته که با نمونه‌های خوشه‌های دیگر تفاوت دارند. حال خوشه‌بندی در تحلیل ارتباط با مشتری چه نتایجی در اختیار ما قرار می‌دهند؟ خوشه‌بندی اطلاعات مفیدی از جمله جدول که درصد تعلق هر گروه و تعداد مشتریان را در گروه‌های مشابه مشخص می‌کند. مشاهده بعدی مشخص می‌شود که با توجه به ویژگی‌های که دارد در کدام خوشه قرار می‌گیرد و همچین این روش تعیین می‌کند که هر خوشه بر اساس ویژگی چه نوع مشتریانی را شامل می‌شود. از دیگر کاربردهای این روش که برای تحلیل دقیق‌تر همراه با الگوریتم قواعد انجمانی می‌توان عادت خرید هر مشتری را تحلیل کرد.

مارکس و اسپانسر یکی از معتبرترین خرده فروشی‌های چندملیتی در بریتانیا است. با این حال این کسب و کار در سالهای اخیر دریافت تنها با اعتبار خود نمی‌تواند سودآوری داشته باشد و نیاز به اطلاعات بیشتری از مشتریان خود است. از این رو با خوشه‌بندی، مشتریان را به 11 گروه طبقه‌بندی کرد. بعد از تجزیه و تحلیل فراوان، این شرکت تصمیم گرفت که در هر انبار دقیقاً کالاهایی را نگهداری کند که مشتریان می‌خواهند. در صورتی که تا آن زمان فروشگاه‌ها با توجه به متراد مرتع انبار، کالا را ذخیره می‌کردند. در یک رستوران بزرگ نیز از طریق دوربین‌های مدار بسته اطلاعات صفت را به رایانه ارسال و پردازش هر چند دقیقه یکبار، تشخیص می‌دهند که یک نتایجی روی صفحه نمایش برای مشتریان صفت نمایش داده شود. به عنوان مثال اگر صفت طولانی باشد، گرینه‌های فست فود بیشتر نمایش داده می‌شود تا تعداد بیشتری فروخته شود و اگر صفت کوتاه‌تر باشد صفحه منو غذاهایی را نمایش می‌دهد که قیمت بالاتر و سود بیشتری دارند اما نیاز به زمان بیشتر برای تهیه دارند.

### قواعد پیوند

قواعد پیوند یکی از روش‌های ناراهنماییده است. در بازاریابی برای فروش همراه محصولات در پیوند با قلمی که مشتری در حال بررسی آن است رایج است. در این روش هدف شناسایی خوشه‌هایی از اقلام در پایگاه داده‌های تراکنش‌گونه است. و معمولاً در این روش با سوالات زیر روبرو هستیم

کدام گروه‌های محصولات، گرایش به خریداری شدن با هم را دارند؟ ►

چه چیزی با چه چیزی فروش می‌رود؟ ►

### پالایش گروهی

یک روش ناراهنماییده است. هدف فراهم کردن توصیه‌های شخصی‌سازی‌شده‌ای است که اطلاعات در سطح کاربر را بالا ببرد. پالایش گروهی، کاربرمبا با یک کاربر شروع می‌کند و سپس کاربرانی را می‌یابد که یک مجموعه قلم مشابه خریده‌اند یا اقلام را به طریق مشابهی رتبه‌بندی کرده‌اند، یک توصیه برای کاربر اولیه بر مبنای آنچه خریداران مشابه خریداری کرده‌اند یا دوست داشته‌اند می‌سازد. به عنوان مثال یک فروشگاه که لوازم جانبی گوشی تلفن همراه می‌فروشد، یک تبلیغ روی قاب گوشی انجام داده است.

## فصل 3 : کاربرد داده کاوی در پژوهش بازار

### مقدمه

یکی از مهم ترین کارهایی که در پژوهش بازار اهمیت دارد، مشخص کردن طبقه اقتصادی-اجتماعی است. در این فصل می خواهیم به کمک داده کاوی به پیش بینی رده بندی های درآمد خانوارهای شهری بر اساس طرح هزینه درآمد سال 1398 که شامل بیست هزار نمونه از تمام خانوارهای شهری ایران است بپردازیم. داده های این گزارش توسط مرکز آمار ایران جمع آوری شده است. تعداد کل نمونه در کلان شهرهای تهران، مشهد، اصفهان، تبریز، شیراز تعداد 2287 خانوار است. می خواهیم به کمک روش های معرفی شده در فصل پیش یعنی درخت رگرسیون، ماشین بردار پشتیبانی<sup>۱۴</sup>، روش k نزدیک ترین همسایه، شبکه عصبی و رگرسیون ترتیبی به رده بندی طبقه اقتصادی اجتماعی از نمونه گرفته شده بپردازیم. در هر بخش داده ها را به دو بخش داده های آموزشی<sup>۱۵</sup> و داده های اعتبارسنجی<sup>۱۶</sup> تقسیم کرده که 50 درصد داده ها را به داده آموزشی اختصاص می دهیم و 35 درصد را نیز به بخش اعتبارسنجی و 15 درصد مابقی را به بخش آزمون اختصاص می دهیم. در ابتدا مدل را بر روی داده آموزشی برآذش داده و بار دیگر برای ارزیابی بیشتر کفایت مدل، مدل حاصله در بخش آموزشی را به داده های اعتبار سنجی برآذش می دهیم، اگر میزان درستی در داده های اعتبارسنجی خیلی نسبت به داده های آموزشی کاهش پیدا کند مدل حاصله بیش برآذش دارد. و در نهایت با بررسی و مقایسه میزان درستی هر روش پرداخته و در نهایت بهترین مدل را انتخاب کنیم.

---

Support vector machines<sup>۱۴</sup>

Training data<sup>۱۵</sup>

Validation data<sup>۱۶</sup>

## متغیرها

### الف) متغیرها پیشگو

در این بخش به معرفی 68 متغیر استفاده شده در این طرح می‌برداریم و رده‌های هریک را مشخص کردیم

ردیف	نام متغیر	تعریف متغیر
1	Address	کد یازده رقمی برای شناسایی خانوار
2	Ostan	کد استان
3	Tedad	تعداد اعضای خانوار
4	Gender	جنسیت سرپرست خانوار دارای دو رده ی مرد=1 و زن=2
5	Age	سن سرپرست خانوار
6	Savad	میزان سواد سرپرست خانوار دارای دو رده ی با سواد=1 و بی سواد=2
7	InEdu	سرپرست خانوار تحصیل می کند؟ دارای دو رده ی بله=1 و خیر=2
8	Madrak	مدرک تحصیلی سرپرست خانوار دارای 9 رده ی ابتدایی/سواد آموزی=1، راهنمایی/متوسطه=2، متوسطه/متوسطه=3، دیپلم و پیش دانشگاهی=4، فوق دیپلم/کاردانی=5، لیسانس/کارشناسی=6، کارشناسی ارشد و دکترای حرفه ای=7، دکترای تخصصی=8، سایر و غیر رسمی=9
9	Faaliat	وضعیت فعالیت سرپرست خانوار دارای 6 رده ی شاغل=1، بیکار(جویای کار)=2، دارای درآمد بدون کار=3، محصل=4، خانه دار=5، سایر=6
10	T.shaghel	تعداد افراد شاغل در خانوار
11	T.M.S	نحوه تصرف منزل مسکونی دارای 6 رده ی ملکی عرصه و اعیان=1، ملکی اعیان=2، اجاری=3، رهن=4، در برابر خدمت=5، رایگان=6
12	T.O	تعداد اتاق در اختیار

سطح زیر بنای محل سکونت	S.Z	13
نوع اسکلت بنای محل سکونت دارای سه رده‌ی فلزی=1، بتن آرمه=2، سایر=3	N.S	14
مصالح عمده بنای محل سکونت دارای 8 رده‌ی آجر و آهن یا سنگ و آهن=1، آجر و چوب یا سنگ و چوب=2، بلوک سیمانی(با هر نوع سقف)=3، تمام آجر یا سنگ و آجر=4، تمام چوب=5، خشت=6، سایر=7	Masleh	15
خانوار از اتومبیل شخصی استفاده میکند؟ خیر=0 و بله=1	oto	16
خانوار از موتورسیکلت استفاده میکند؟ خیر=0 و بله=1	motor	17
خانوار از دوچرخه استفاده میکند؟ خیر=0 و بله=1	do	18
خانوار از رادیو استفاده میکند؟ خیر=0 و بله=1	radio	19
خانوار از ضبط استفاده میکند؟ خیر=0 و بله=1	zabt	20
خانوار از تلویزیون سیاه و سفید استفاده میکند؟ خیر=0 و بله=1	TV.s	21
خانوار از تلویزیون رنگی استفاده میکند؟ خیر=0 و بله=1	TV.r	22
خانوار از انواع ویدئو، VCD و DVD استفاده میکند؟ خیر=0 و بله=1	DVD	23
خانوار از انواع یارانه و تبلت استفاده میکند؟ خیر=0 و بله=1	pc	24
خانوار از تلفن همراه استفاده میکند؟ خیر=0 و بله=1	mobile	25
خانوار از فریزر استفاده میکند؟ خیر=0 و بله=1	freeizer	26
خانوار از یخچال استفاده میکند؟ خیر=0 و بله=1	yakhchal	27
خانوار از یخچال فریزر استفاده میکند؟ خیر=0 و بله=1	yakhchal.f	28
خانوار از اجاق گاز استفاده میکند؟ خیر=0 و بله=1	gaz	29
خانوار از جارو برقی استفاده میکند؟ خیر=0 و بله=1	jaro.b	30
خانوار از ماشین لباسشویی استفاده میکند؟ خیر=0 و بله=1	m.lebas	31
خانوار از چرخ خیاطی استفاده میکند؟ خیر=0 و بله=1	charkh.kh	32
خانوار از پنکه استفاده میکند؟ خیر=0 و بله=1	panke	33
خانوار از کولر آبی متحرک استفاده میکند؟ خیر=0 و بله=1	cooler.a	34

خانوار از کولر گازی متحرک استفاده میکند؟ خیر=0 و بله=1	cooler.g	35
خانوار از ماشین ظرفشویی استفاده میکند؟ خیر=0 و بله=1	m.zarf	36
خانوار از مایکروویو و انواع فرهای هالوژن دار استفاده میکند؟ خیر=0 و بله=1	microfer	37
خانوار از آب لوله کشی استفاده میکند؟ خیر=0 و بله=1	ab.l	38
خانوار از برق استفاده میکند؟ خیر=0 و بله=1	bargh	39
خانوار از گاز لوله کشی استفاده میکند؟ خیر=0 و بله=1	gaz.l	40
خانوار از تلفن ثابت استفاده میکند؟ خیر=0 و بله=1	tel	41
خانوار از دسترسی به اینترنت استفاده میکند؟ خیر=0 و بله=1	internet	42
خانوار از حمام استفاده میکند؟ خیر=0 و بله=1	hamam	43
خانوار از آشپزخانه استفاده میکند؟ خیر=0 و بله=1	ashpazkhane	44
خانوار از کولر آبی ثابت استفاده میکند؟ خیر=0 و بله=1	cooler.a.s	45
خانوار از برودت مرکزی استفاده میکند؟ خیر=0 و بله=1	broodat.m	46
خانوار از حرارت مرکزی استفاده میکند؟ خیر=0 و بله=1	hararat.m	47
خانوار از پکیج استفاده میکند؟ خیر=0 و بله=1	package	48
خانوار از کولر گازی ثابت استفاده میکند؟ خیر=0 و بله=1	cooler.g.s	49
خانوار از شبکه عمومی فاضلاب استفاده میکند؟ خیر=0 و بله=1	fazelab	50
نوع سوخت برای پخت و پز دارای ده رده نفت سفید=1، گازوییل=2، گاز مایع=3، گاز طبیعی=4، برق=5، هیزم و زغال=6، سوخت حیوانی=7، زغال سنگ=8، سایر سوخت ها=9، هیچکدام=10	sookht.p	51
نوع سوخت برای ایجاد گرما دارای ده رده نفت سفید=11، گازوییل=12، گاز مایع=13، گاز طبیعی=14، برق=15، هیزم و زغال=16، سوخت حیوانی=17، زغال سنگ=18، سایر سوخت ها=19، هیچکدام=20	sookht.g	52

نوع سوخت برای تهیه آب گرم دارای ده رده نفت سفید=21، گازوییل=22، گاز مایع=23، گاز طبیعی=24، برق=25، هیزم و زغال=26، سوخت حیوانی=27، زغال سنگ=28، سایر سوخت ها=29، هیچکدام=30	sookht.ab	53
هزینه های خوراکی و دخانیات خانوار در یکماه گذشته به ریال	Hazine_Khorakivadokhani	54
هزینه های نوشیدنی خانوار در یکماه گذشته به ریال	Hazine_Noshidani	55
هزینه پوشак خانوار در یکماه گذشته به ریال	Hazine_Pushak	56
هزینه های مسکن خانوار در یکماه گذشته به ریال	Hazine_Maskan	57
هزینه های مبلمان خانوار در یکماه گذشته به ریال	Hazine_Mobleman	58
هزینه های بهداشت خانوار در یکماه گذشته به ریال	Hazine_Behdasht	59
هزینه های حمل و نقل خانوار در یکماه گذشته به ریال	Hazine_Hamlonaghl	60
هزینه های ارتباطات خانوار در یکماه گذشته به ریال	Hazine_Ertebatat	61
هزینه های تفریحات خانوار در یکماه گذشته به ریال	Hazine_Tafrihat	62
هزینه های غذای آماده خانوار در یکماه گذشته به ریال	Hazine_Ghazayeamade	63
هزینه های کالا و خدمات خانوار در یکماه گذشته به ریال	Hazine_kalavakhadamat	64
درآمد مزد خانوار در 12 ماه گذشته به ریال	Daramad_Mozd	65
درآمد آزاد خانوار در 12 ماه گذشته به ریال	Daramad_Azad	66
درآمدهای متفرقه خانوار در 12 ماه گذشته به ریال	Daramad_Motefaraghe	67
مبلغ دریافتی یارانه خانوار در 12 ماه گذشته به ریال	Daramad_Yarane	68

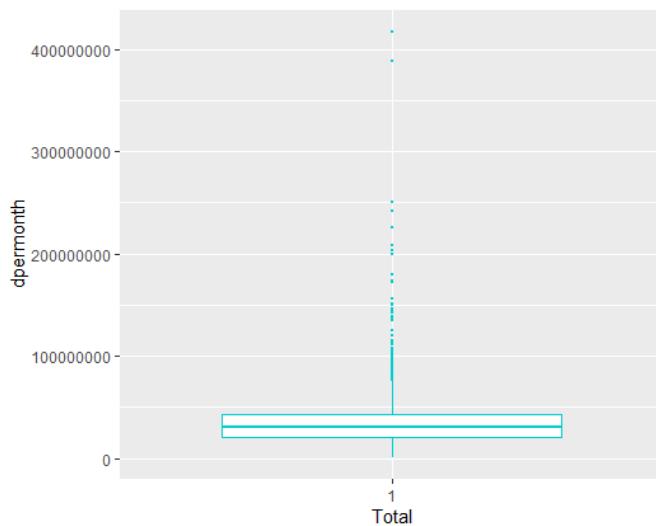
طبقه اقتصادی و اجتماعی یکی از معیارهای اساسی در پژوهش بازار مقایسه وضعیت آگاهی و مصرف در طبقات اقتصادی اجتماعی است. در این پژوهه این معیار با حاصل جمع کل درآمد که شامل مجموع درآمد مزد، آزاد، متفرقه و یارانه خانوار است و نتیجه حاصل را با تقسیم بر تعداد ماههای سال درآمد ماهانه خانوار (dpermonth) حاصل می‌شود

**summary(dpermonth)**

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
455000	20038334	30000000	35171575	42482500	417576667

درآمد کمتر از چارک اول را سطح D در نظر گرفتیم و بین چارک اول و دوم سطح C و از چارک دوم تا سوم

سطح B و بیشتر از چارک سوم را سطح A در نظر گرفتیم که نمودار آن در صفحه بعد مشخص است:



cat	Freq	class
D	572	under 2million toman
C	589	2-3 milion toman
B	554	3-4 milion toman
A	572	more than 4million toman

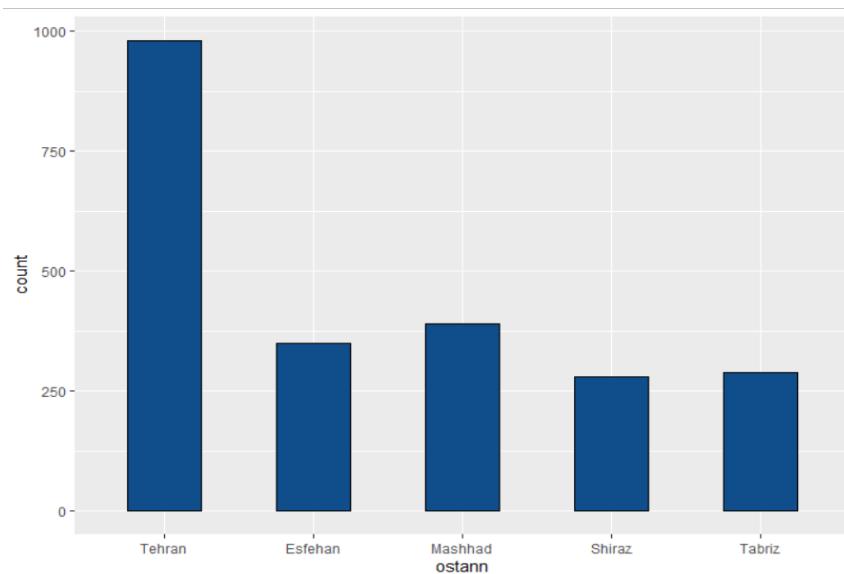
نمودار جعبه‌ای درآمد کل

در box-plot فوق مشاهده می‌کنید که تعدادی مقادیر دورافتاده در درآمدهای بالا داریم.

## تحلیل اکتشافی

الف) نمودارهای تک متغیره

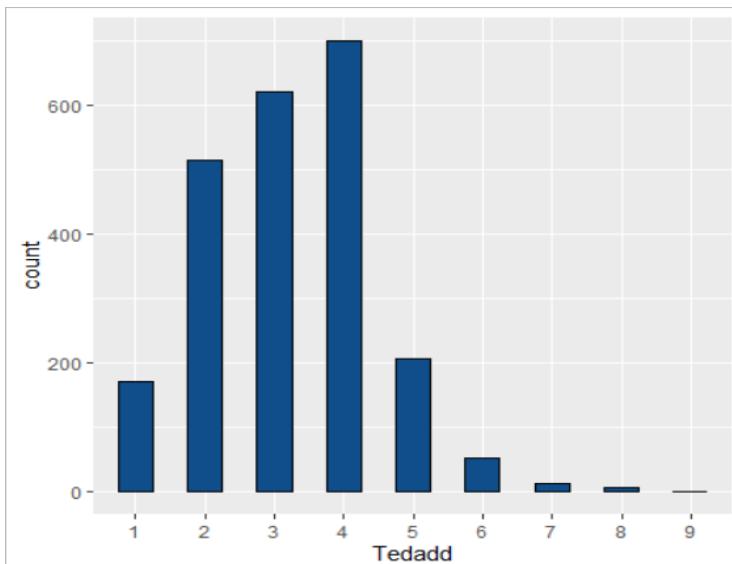
1. شهر



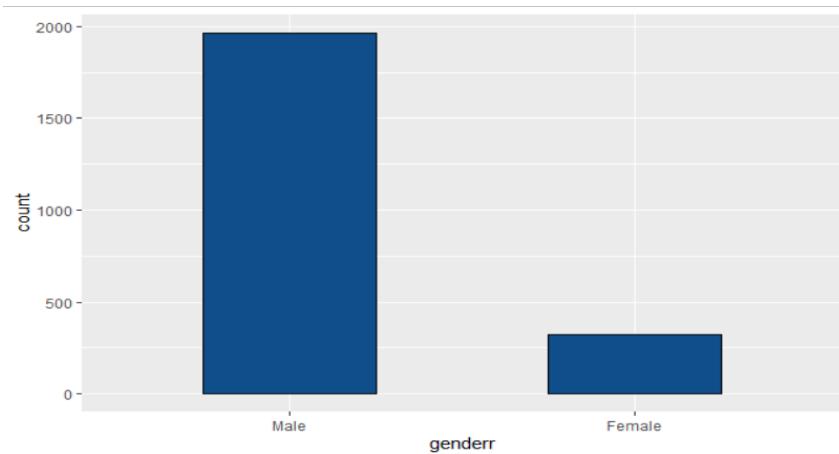
ostann	Freq	perc
Tehran	981	43 %
Esfahan	349	15 %
Mashhad	389	17 %
Shiraz	280	12 %
Tabriz	288	13 %
sum	2287	100 %

شهر تهران با 981 نفر بیشترین تعداد را دارد.

2. تعداد اعضای خانوار

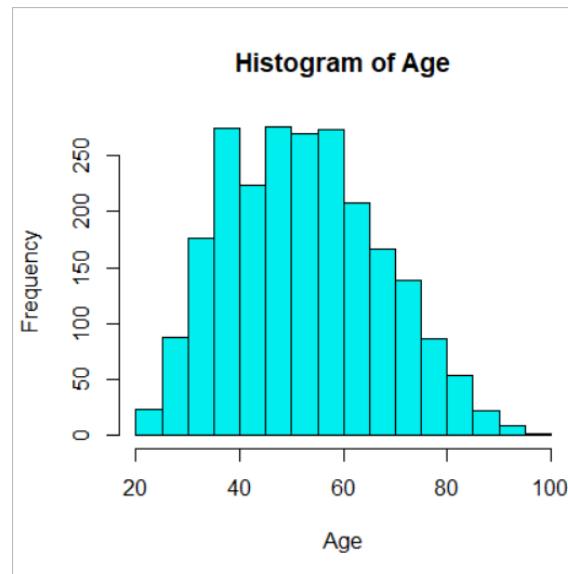


.3 جنسیت سرپرست خانوار

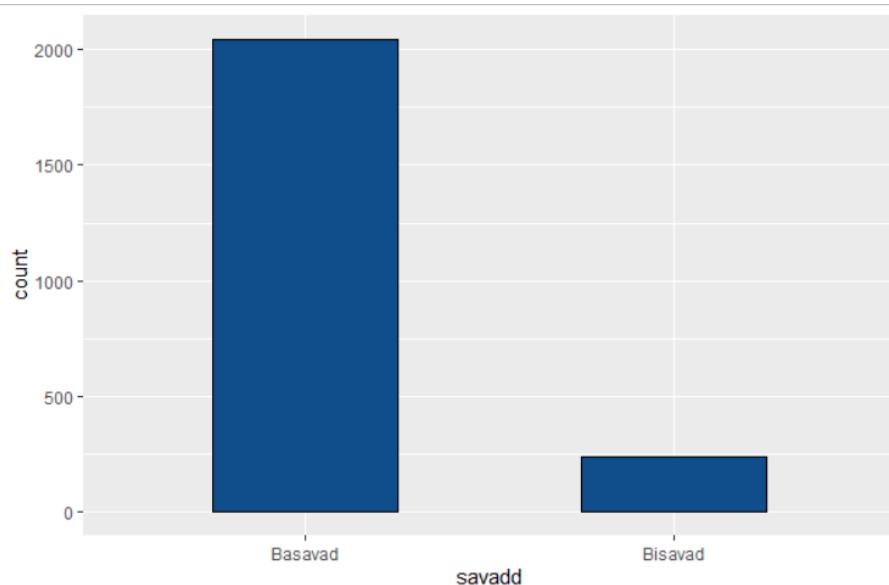


86 درصد از سرپرستان خانوار مرد می‌باشند.

.4 سن سرپرست خانوار



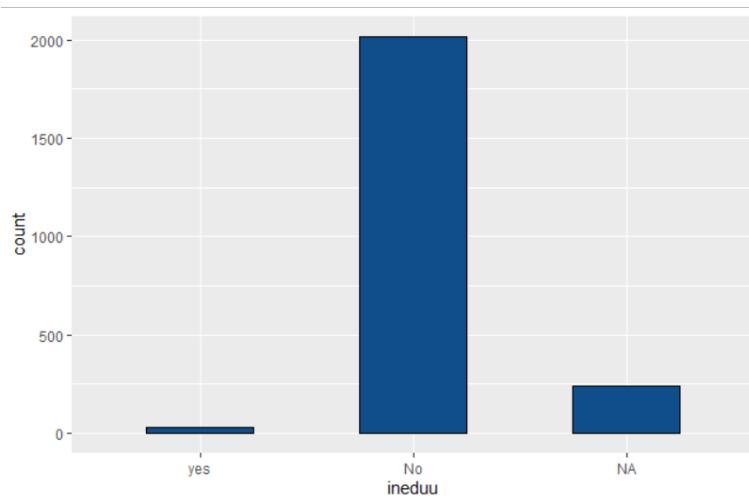
. سواد سرپرست خانوار



savadd	Freq	perc
Basavad	2046	89 %
Bisavad	241	11 %
sum	2287	100 %

89 درصد از سرپرستان خانوار با سواد هستند.

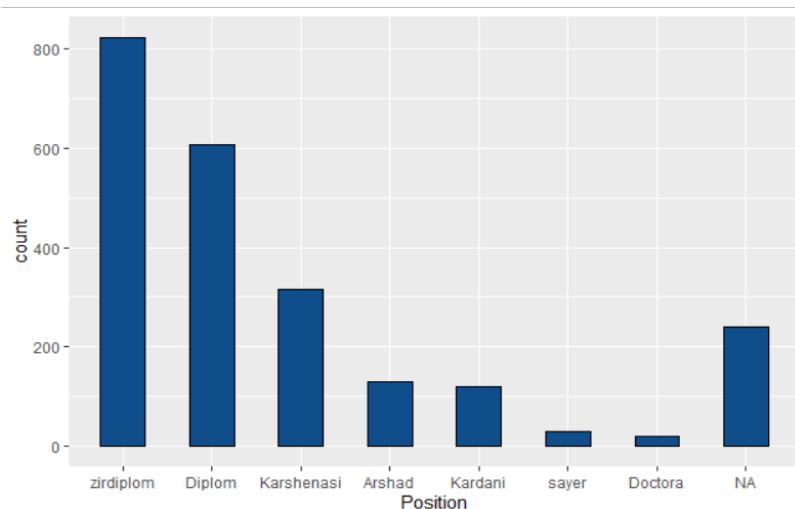
## 6. در حال تحصیل بودن یا نبودن سرپرست خانوار



ineduu	Freq	perc
yes	28	1 %
No	2018	99 %
sum	2046	100 %

تنها 1 درصد سرپرستان خانوار در حال تحصیل هستند و تعداد افرادی که در این نمودار گشوده به حساب آمدند یعنی 241 نفر که بیسواند بودند.

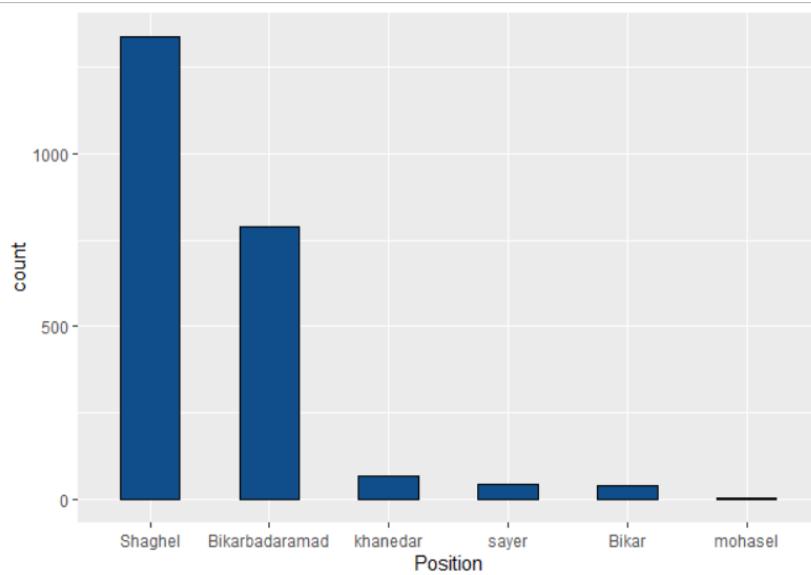
## 7. مدرک تحصیلی سرپرست خانوار



madrakk	Freq	perc
zirdiplom	823	36 %
Diplom	607	27 %
Karshenasi	316	14 %
Arshad	129	6 %
Kardani	121	5 %
sayer	29	1 %
Doctora	21	1 %
sum	2046	89 %

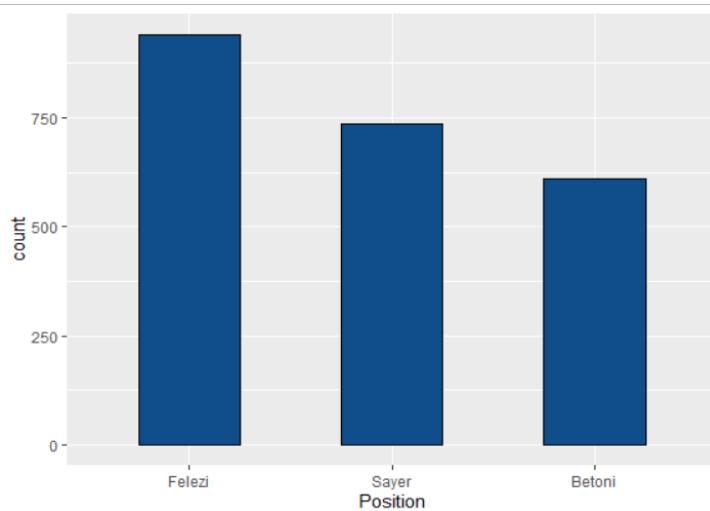
36 درصد سرپرستان خانوار زیر دیپلم هستند.

## 8. فعالیت سرپرست خانوار



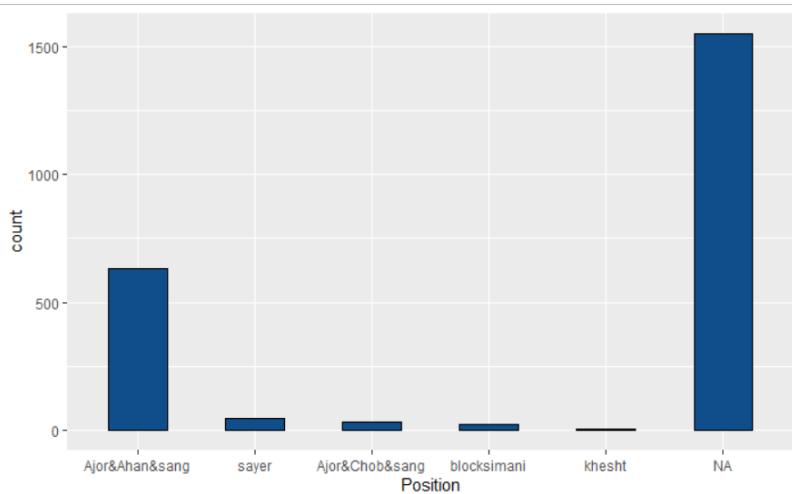
تقريبا 60 درصد سرپستان خانوار شاغل هستند

#### 9. نوع اسکلت ساختمان محل سکونت



بیشترین فرآوی نوع اسکلت مربوط به اسکلت فلزی است.

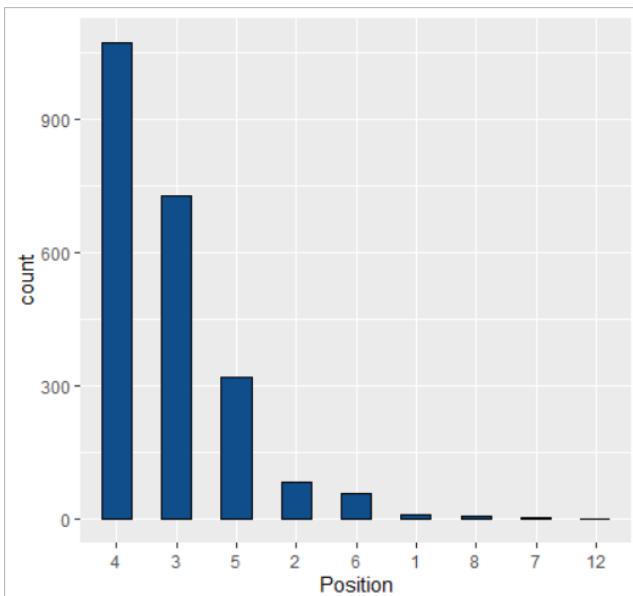
## 10. نوع مصالح ساختمان محل سکونت



Maslehh	Freq	perc
Ajor&Ahan&sang	630	86 %
sayer	45	6 %
Ajor&Chob&sang	31	4 %
blocksimani	23	3 %
khesht	7	1 %
Ajor	0	0 %
Chob	0	0 %
sum	736	100 %

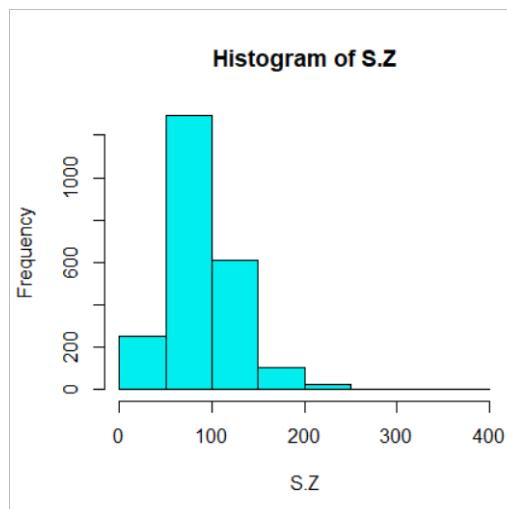
درصد افراد جنس مصالح استفاده شده در ساختمان آجر و آهن و سنگ است. 1551 داده گمشده داریم و چون تعداد مقادیر گمشده خیلی زیاد است پس این متغیر کاندید حذف است.

## 11. تعداد اتاق محل سکونت

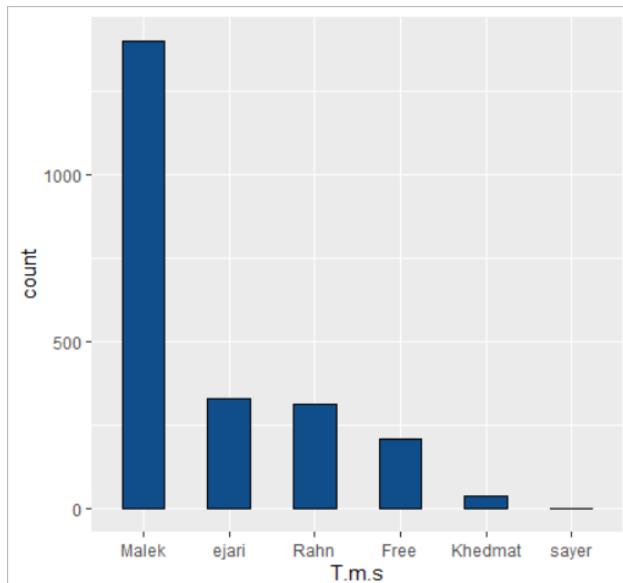


T.O.	Freq	perc
4	1075	47 %
3	730	32 %
5	319	14 %
2	83	4 %
6	58	3 %
1	10	0 %
8	6	0 %
7	5	0 %
12	1	0 %
sum	2287	100 %

12. سطح زیربنای ساختمان محل سکونت



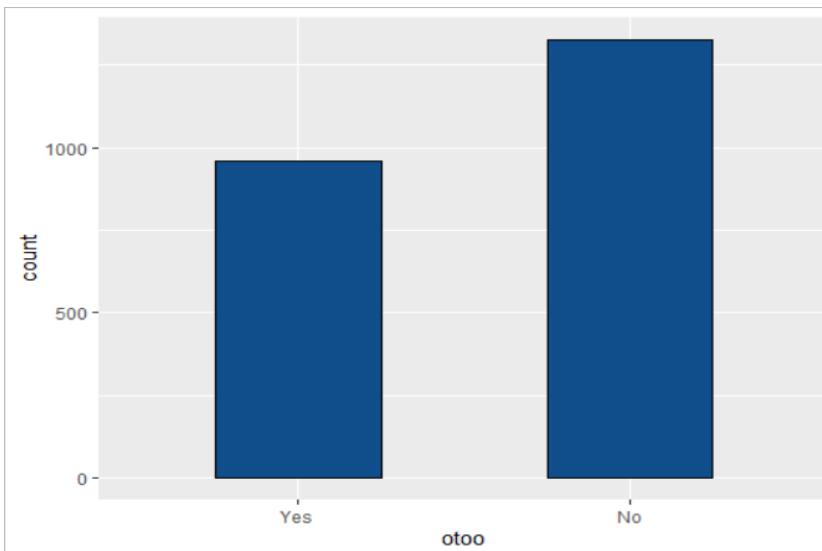
13. نحوه تصرف ساختمان محل سکونت



T.M.S.	Freq	perc
Malek	1400	61 %
ejari	329	14 %
Rahn	312	14 %
Khedmat	36	2 %
Free	208	9 %
sayer	2	0 %
sum	2287	100 %

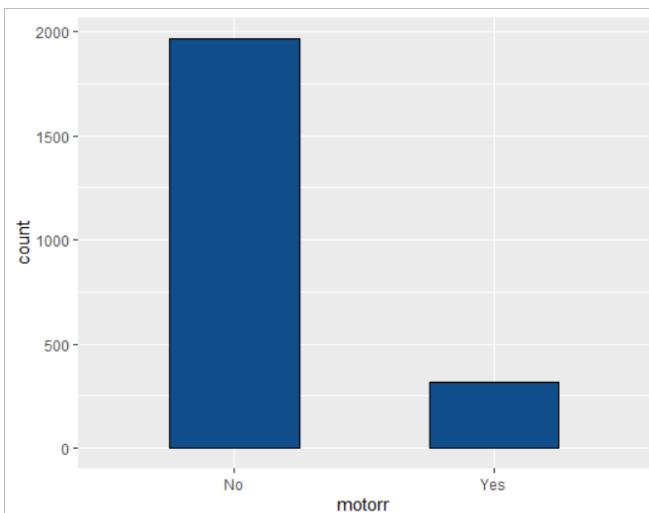
61 درصد سرپرستان خانوار مالک محل سکونت هستند.

14. اتومبیل



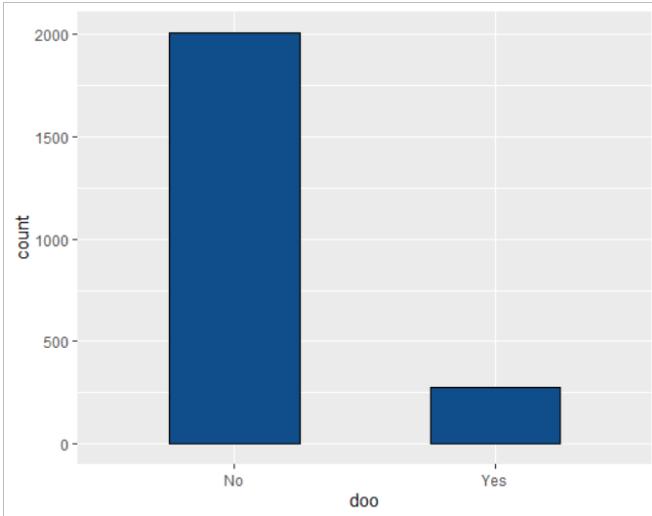
42 درصد سرپرستان خانوار صاحب اتومبیل هستند

15. موتور سیکلت



14 درصد سرپرستان خانوار صاحب موتورسیکلت هستند.

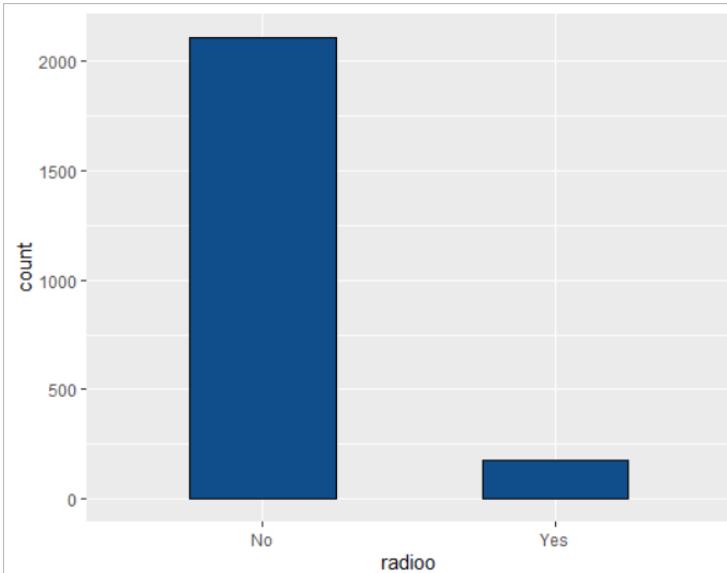
16. دوچرخه



doo	Freq	perc
No	2010	88 %
Yes	277	12 %
sum	2287	100 %

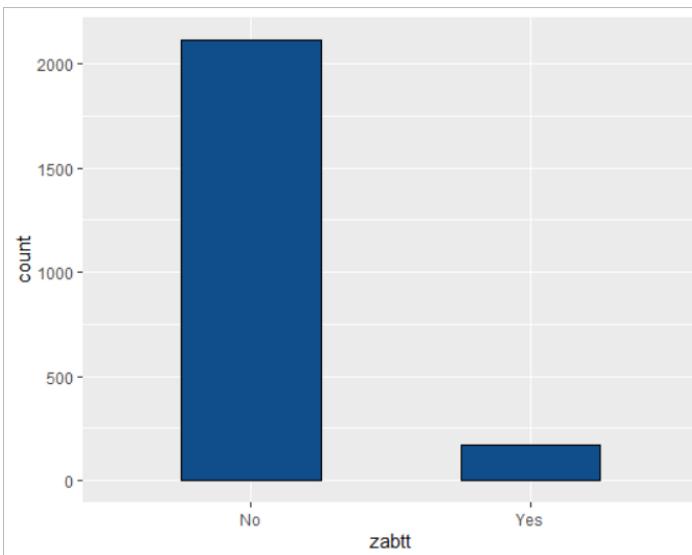
12 درصد سرپرستان خانوار صاحب دوچرخه هستند

.رادیو 17



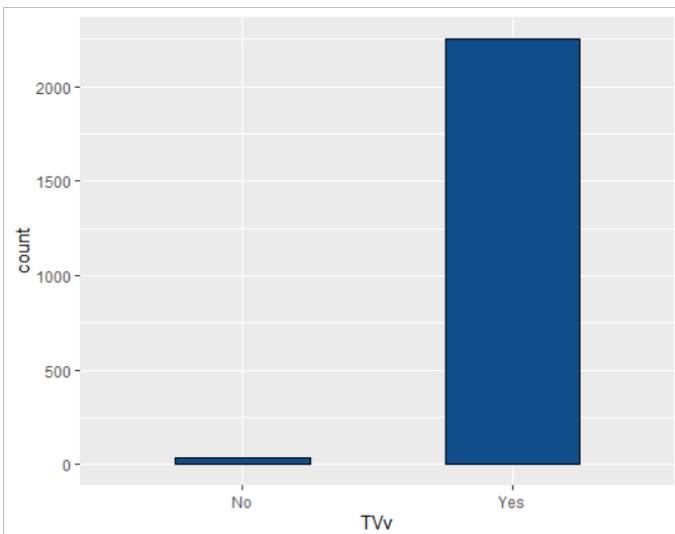
radioo	Freq	perc
No	2110	92 %
Yes	177	8 %
sum	2287	100 %

18. ضبط



zabtt	Freq	perc
No	2116	93 %
Yes	171	7 %
sum	2287	100 %

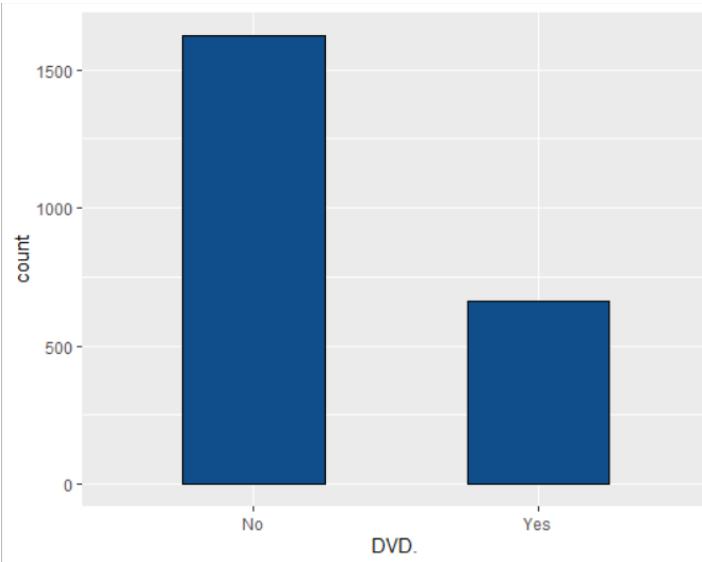
19. تلویزیون



TVv	Freq	perc
No	32	1 %
Yes	2255	99 %
sum	2287	100 %

99 درصد سرپرستان خانوار صاحب تلویزیون هستند. پس این متغیر نیز کاندید حذف است.

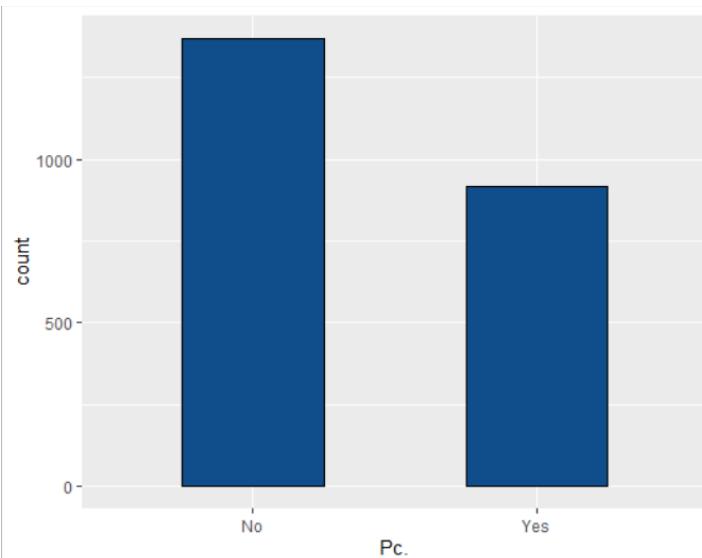
DVD .20



DVD.	Freq	perc
No	1625	71 %
Yes	662	29 %
sum	2287	100 %

امروزه می‌دانیم داشتن یا نداشتن DVD تاثیری چندانی بر سطح درآمدی فرد ندارد.

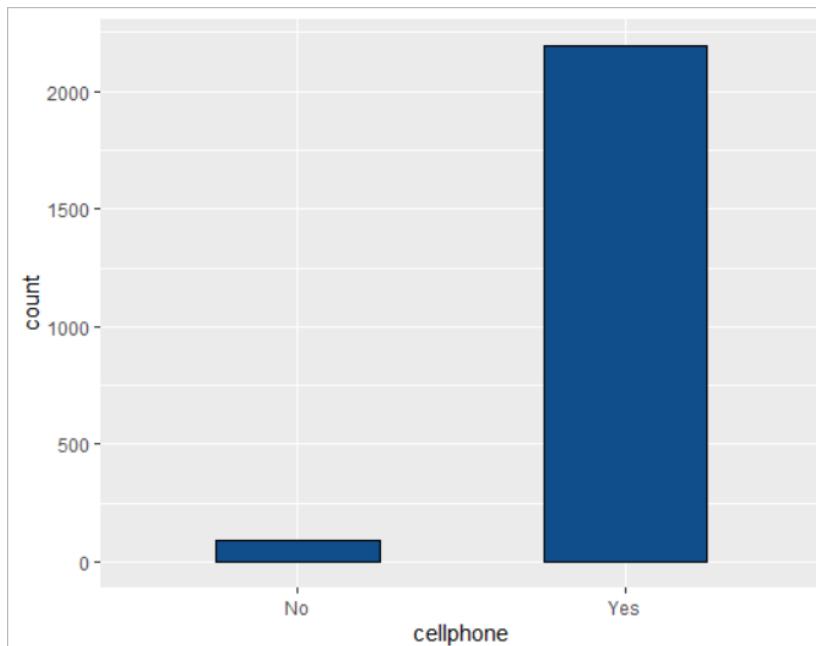
21. کامپیوتر خانگی



Pc.	Freq	perc
No	1370	60 %
Yes	917	40 %
sum	2287	100 %

40 درصد از خانواده‌ها کامپیوتر دارند.

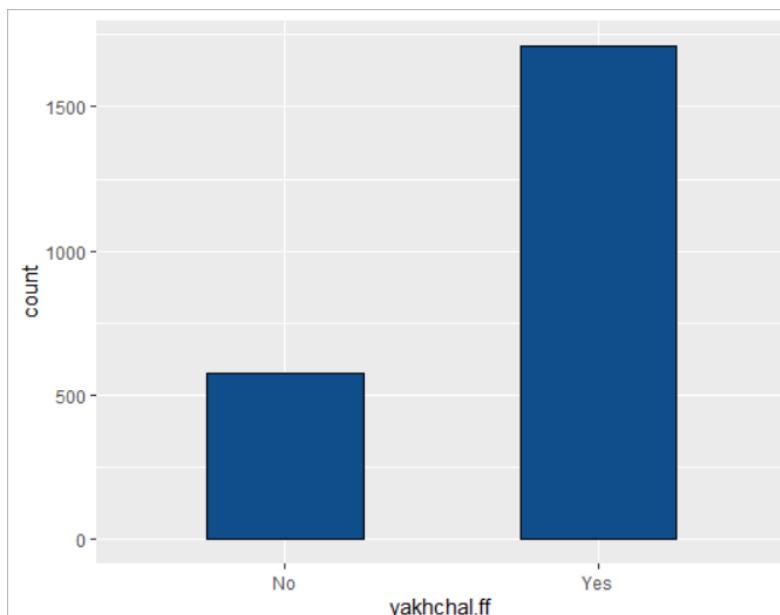
22. تلفن همراه



درصد سرپرستان خانوار تلفن همراه دارند.

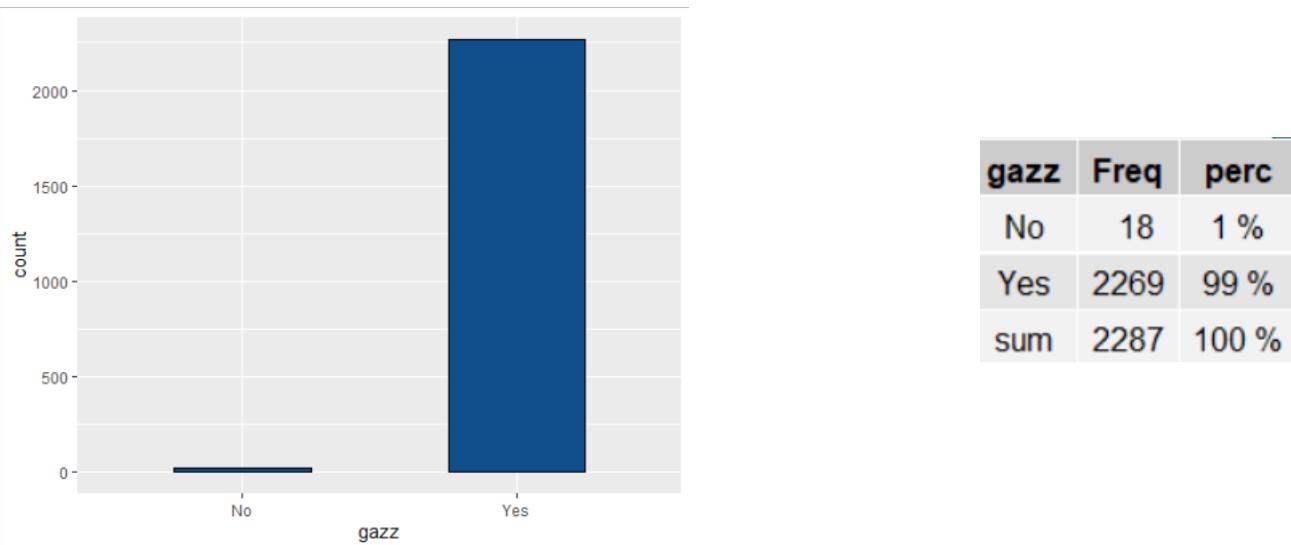
cellphone	Freq	perc
No	92	4 %
Yes	2195	96 %
sum	2287	100 %

23. یخچال فریزر



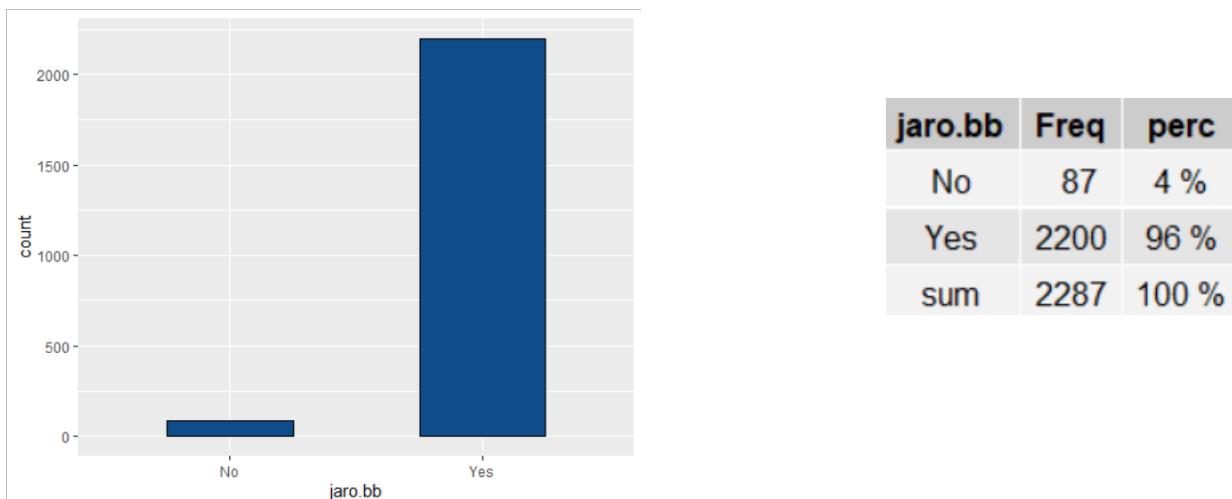
درصد خانواده‌ها صاحب یخچال فریزر هستند

## 24. اjac گاز



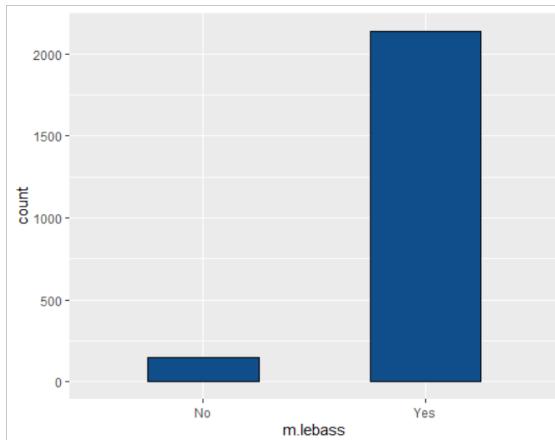
99 درصد خانواده‌ها صاحب اjac گاز هستند. پس این متغیر نیز کاندید حذف است.

## 25. جارو برقی



96 درصد خانواده‌ها صاحب جارو برقی هستند.

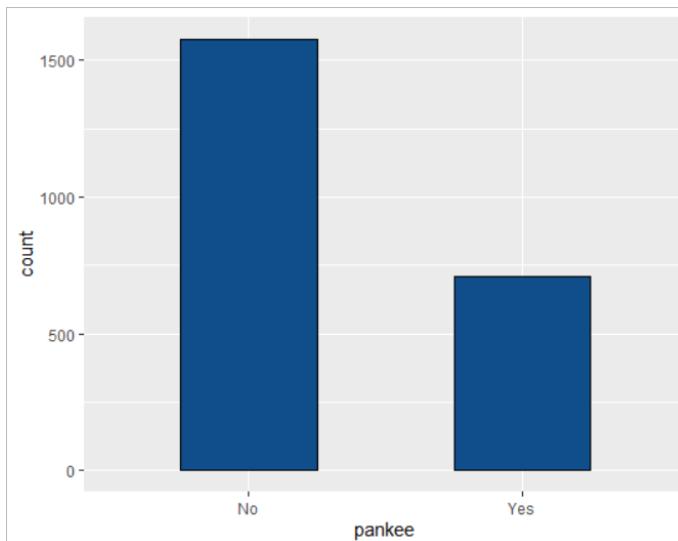
## 26. ماشین لباسشویی



m.lebass	Freq	perc
No	145	6 %
Yes	2142	94 %
sum	2287	100 %

درصد خانواده‌ها ماشین لباس شویی دارند.

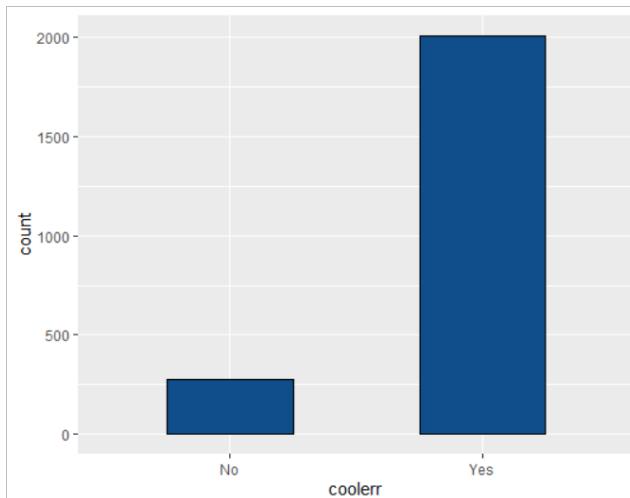
## 27. پنکه



pankee	Freq	perc
No	1579	69 %
Yes	708	31 %
sum	2287	100 %

تقریباً 70 درصد افراد پنکه دارند. اما این متغیر کمکی به هدف مسئله ما نمی‌کند پس کاندید حذف است.

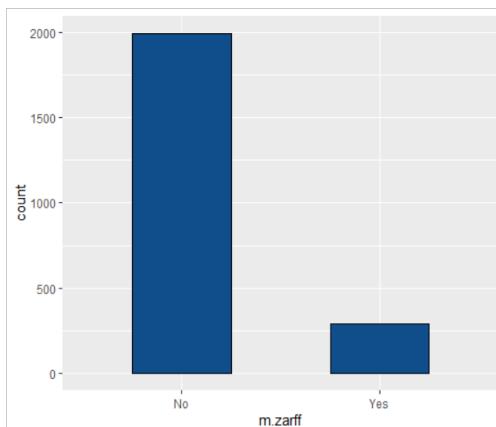
28. کولر



coolerr	Freq	perc
No	278	12 %
Yes	2009	88 %
sum	2287	100 %

درصد افراد صاحب کولر هستند.

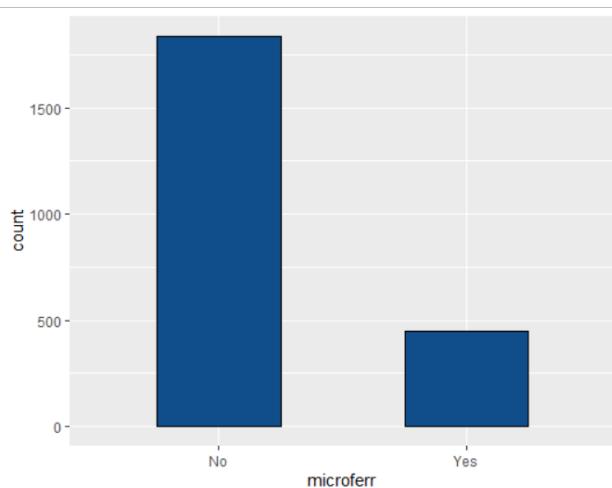
29. ماشین ظرفشویی



m.zarff	Freq	perc
No	1994	87 %
Yes	293	13 %
sum	2287	100 %

درصد خانواده‌ها صاحب ماشین ظرف شویی هستند.

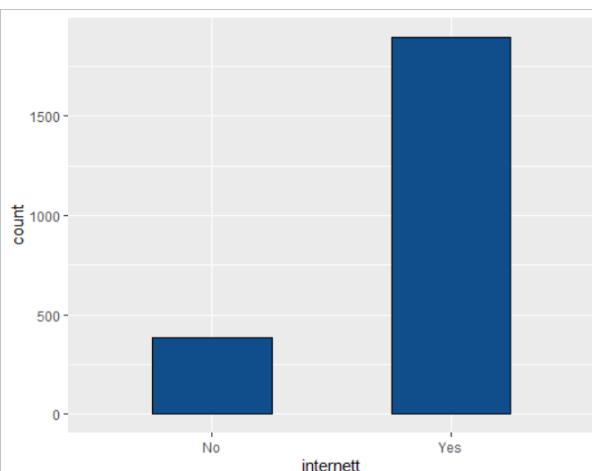
30. مایکروویو



microferr	Freq	perc
No	1838	80 %
Yes	449	20 %
sum	2287	100 %

تنها 20 درصد خانواده‌ها مایکروویو دارند.

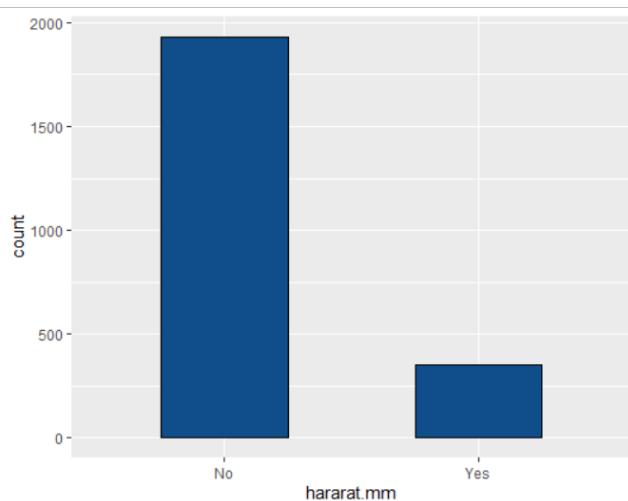
31. اینترنت



internett	Freq	perc
No	387	17 %
Yes	1900	83 %
sum	2287	100 %

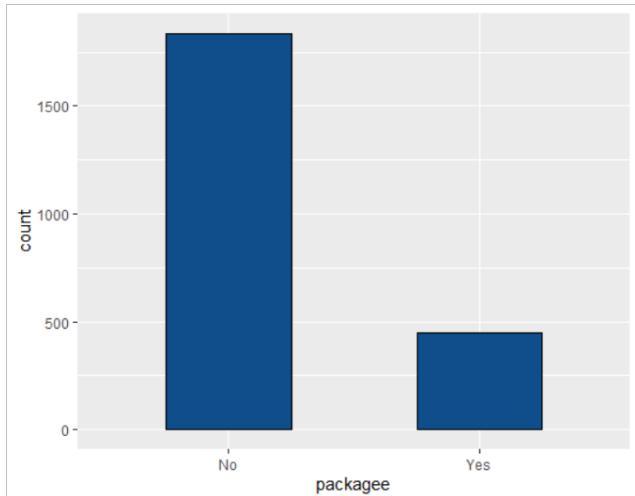
83 درصد خانواده‌ها اینترنت دارند.

32. حرارت مرکزی



hararat.mm	Freq	perc
No	1933	85 %
Yes	354	15 %
sum	2287	100 %

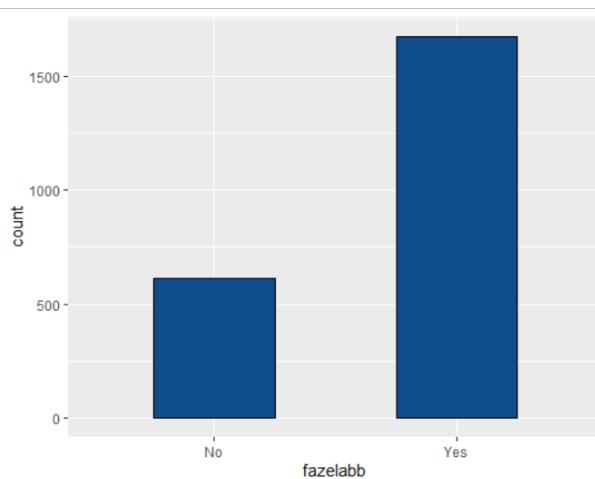
33. پکیج.



packagee	Freq	perc
No	1837	80 %
Yes	450	20 %
sum	2287	100 %

80 درصد خانواده‌ها نمونه پکیج ندارند.

34. فاضلاب



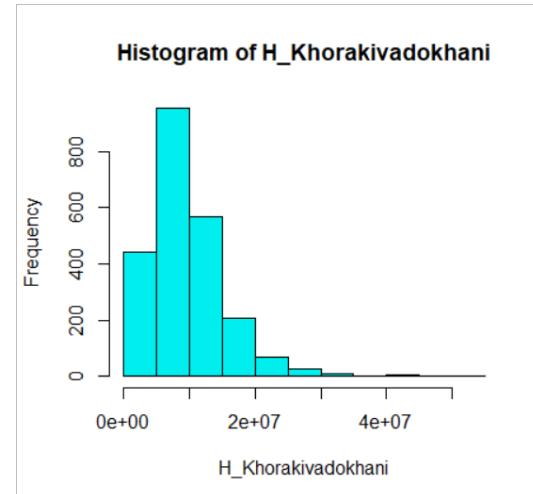
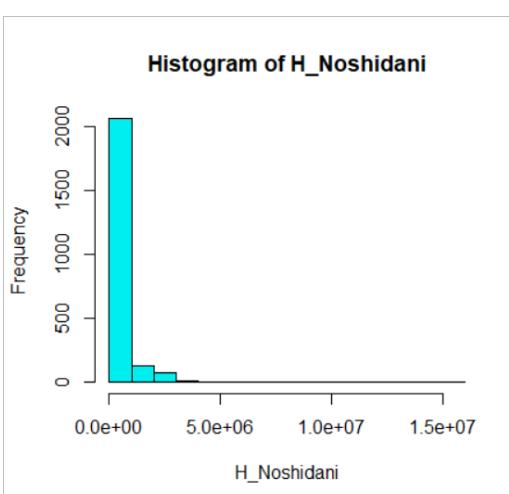
fazelabb	Freq	perc
No	613	27 %
Yes	1674	73 %
sum	2287	100 %

این متغیر چون غالبا تحت کنترل خانواده نیست پس کاندید حذف است

ب) هیستوگرام انوع هزینه و درآمد ماهانه

2) هزینه نوشیدنی

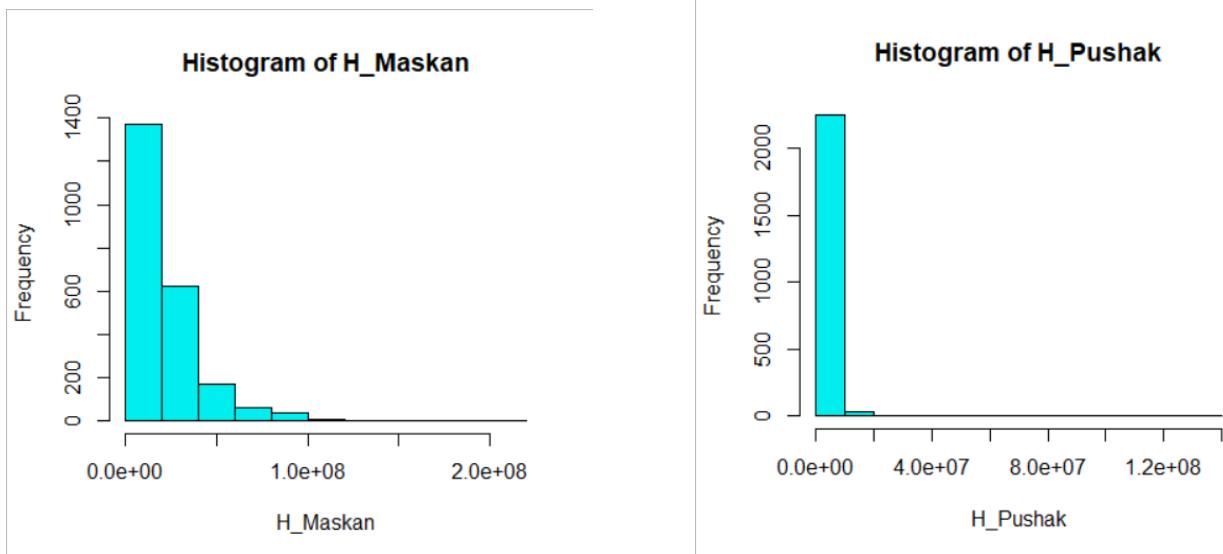
1) هزینه خوراکی دخانیات



هیستوگرام دخانیات خوراکی و دوختایات چوله به راست است.

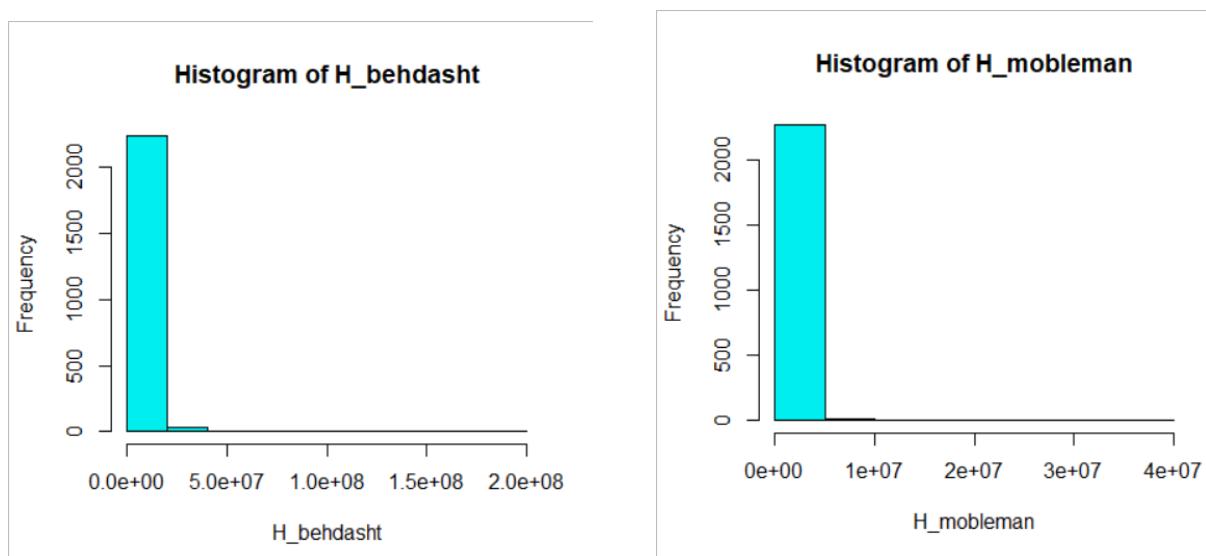
4) هزینه مسکن

3) هزینه پوشاش



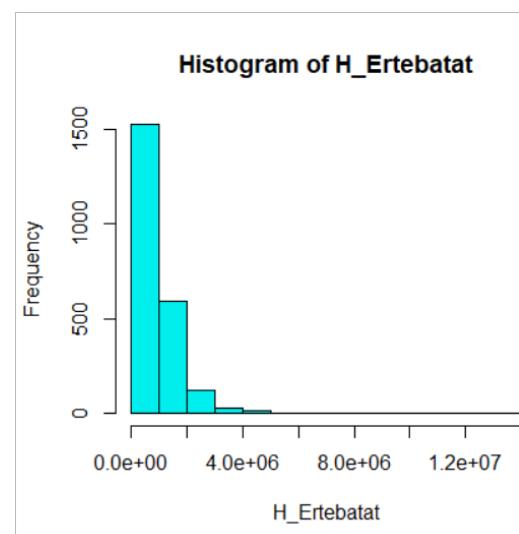
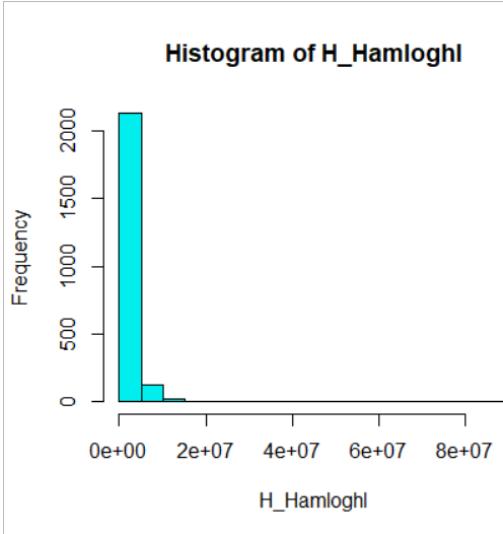
6) هزینه بهداشت

5) هزینه مبلمان



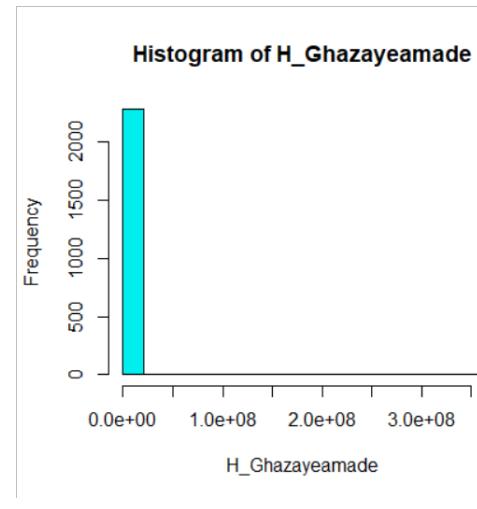
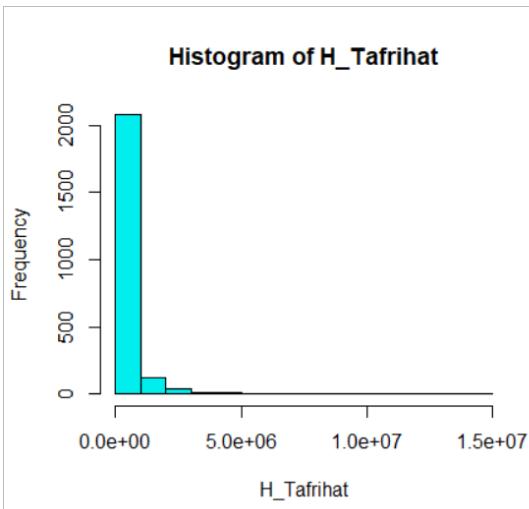
8) هزینه حمل و نقل

7) هزینه ارتباطات

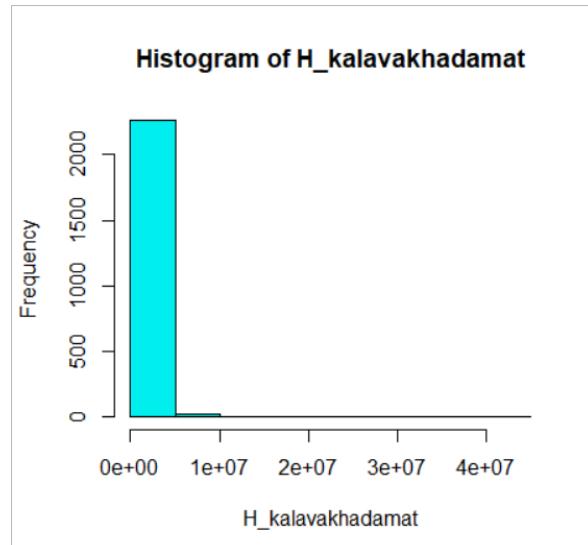


10) هزینه غذای آماده

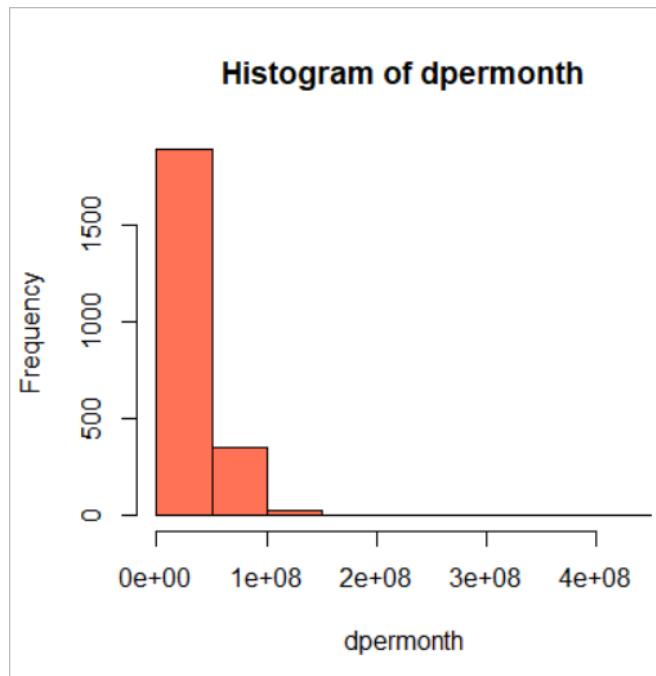
9) هزینه ارتباطات



11) هزینه کالا و خدمات



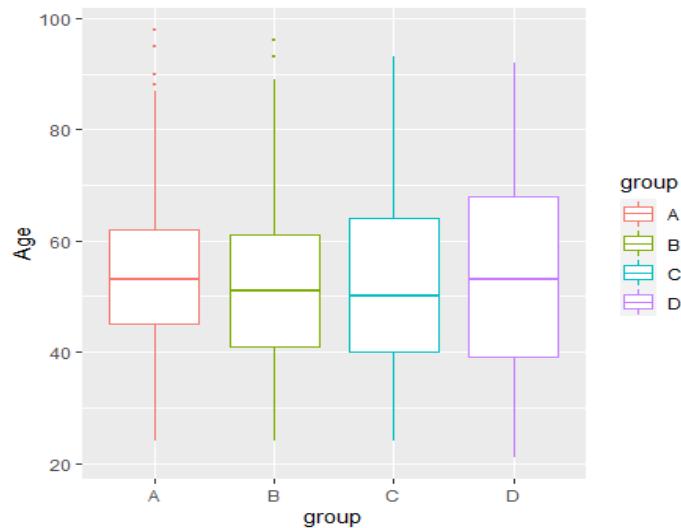
12) متغیر درآمد



### ج) نمودارهای دو متغیره

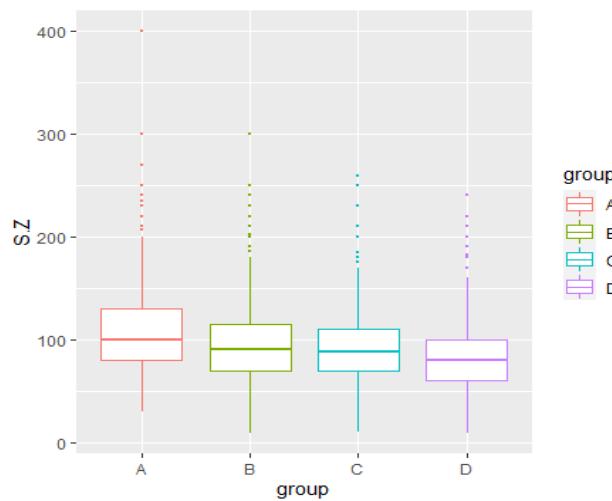
الف) متغیرهای پیشگو و متغیر گروه بندی درآمدی

(1) سن



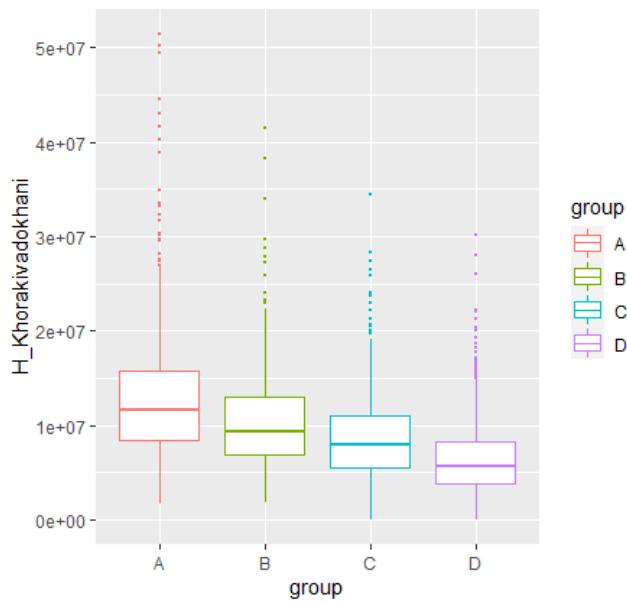
در این نمودار بنظر می آید ارتباط خاصی بین سن و گروه بندی اقتصادی و اجتماعی وجود ندارد

(2) مساحت زیربنای ساختمان محل سکونت



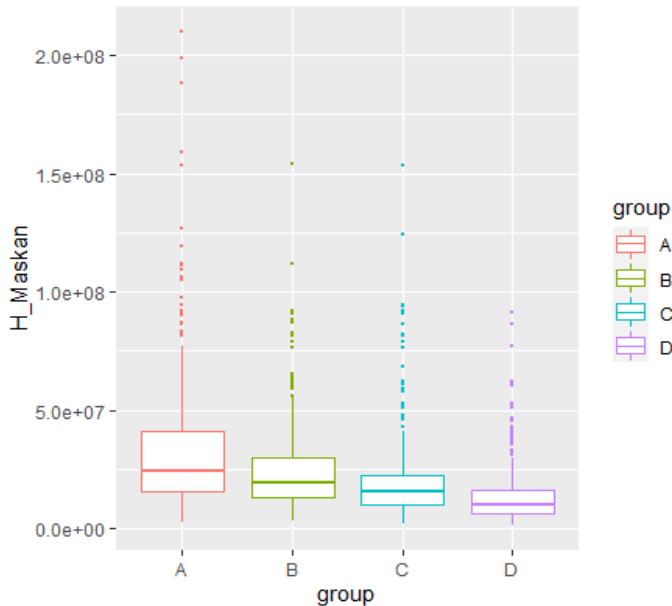
به نظر می آید کسانی که سطح زیربنای بزرگتری دارند در گروه های درآمدی بالاتری قرار دارند

### (3) هزینه خوراکی و دخانیات



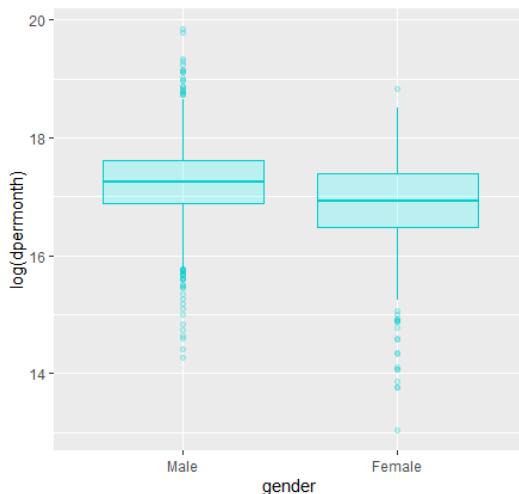
به نظر می‌آید کسانی که هزینه خوراکی و دخانیات بیشتری دارند در گروه‌های درآمدی بالاتری قرار دارند

### (4) هزینه مسکن



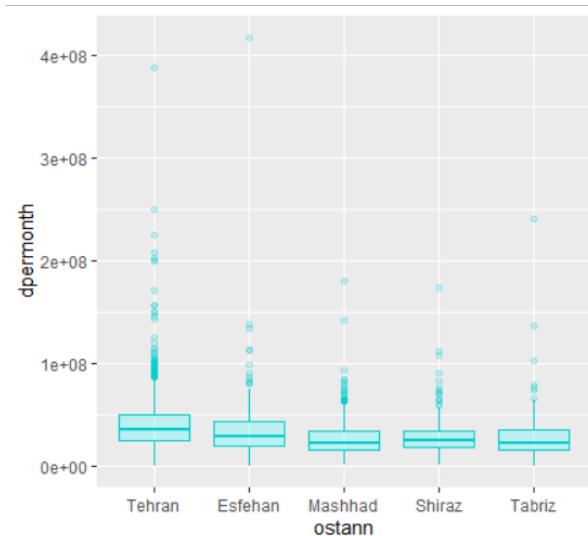
به طور کلی در خانواده‌هایی که در سطح درآمدی بالاتری برخودارن هزینه‌های بیشتری دارند.

(5) جنسیت در مقابل لگاریتم درآمد ماهانه



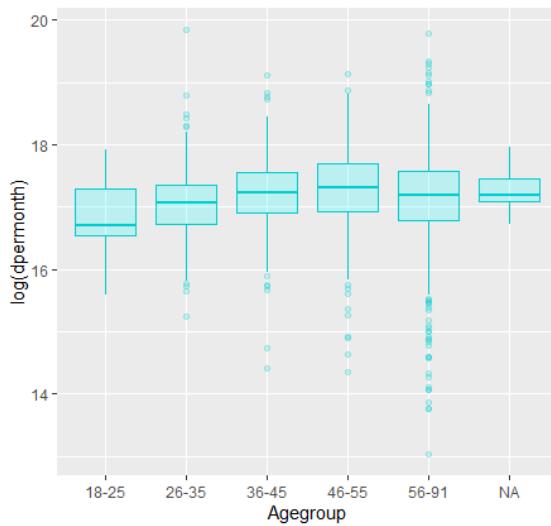
ملاحظه می‌کنید که مردان درآمد بیشتری نسبت به خانم‌ها دارند

(6) شهر محل زندگی و درآمد ماهانه



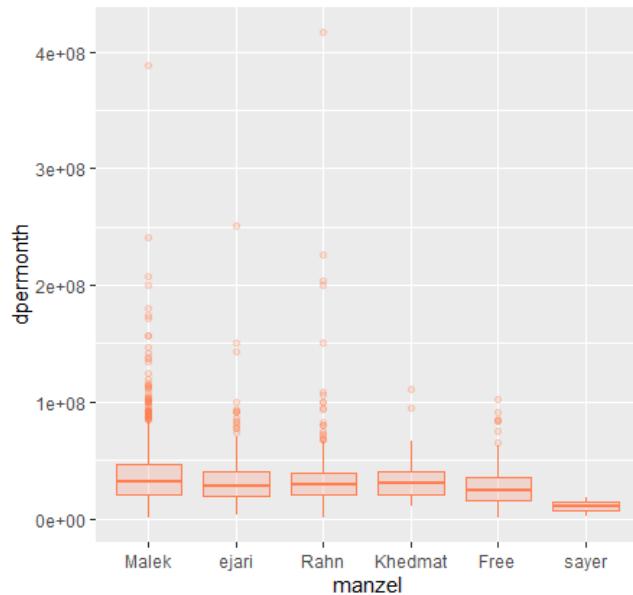
درآمد ماهانه در تهران و اصفهان بیشتر سایر شهرها است

(7) گروه بندی سنی و لگاریتم درآمد



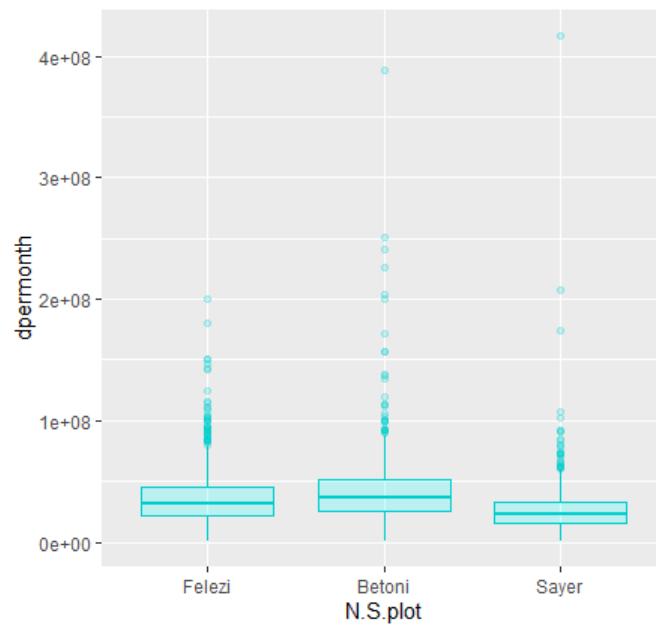
برخلاف نمودار سن و گروههای درآمدی که نشان می‌داد گروههای درآمدی تقریباً از نظر سنی یکسان بودند ولی این نمودار نشان می‌دهد که درامد تا گروه سنی 46–55 سال افزایش میابد و بعد از آن کاهش میابد

#### (8) نوع تصرف ملک و درآمد

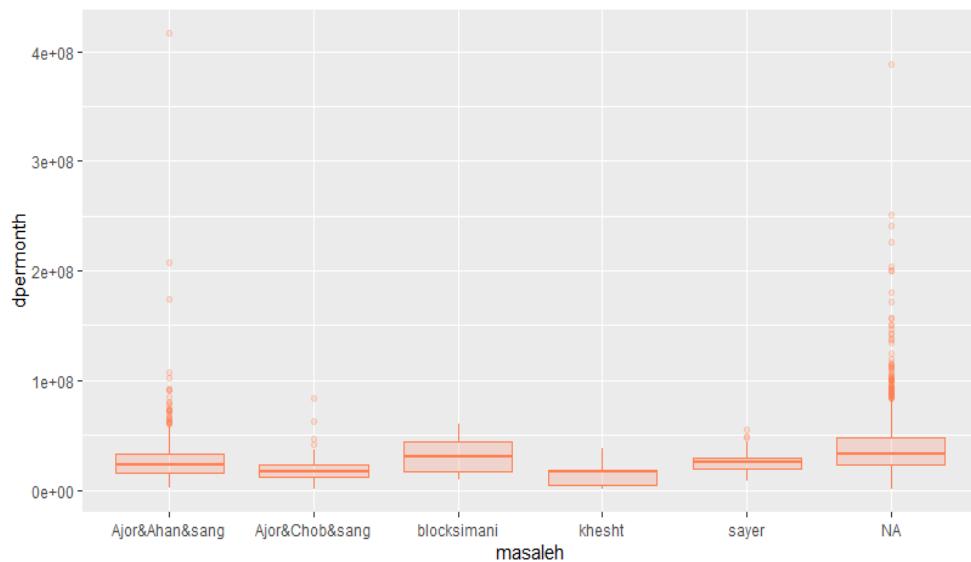


به نظر می‌رسد کسانی که مالک هستند درآمد بیشتری دارند

9) نوع اسکلت ساختمان و درآمد

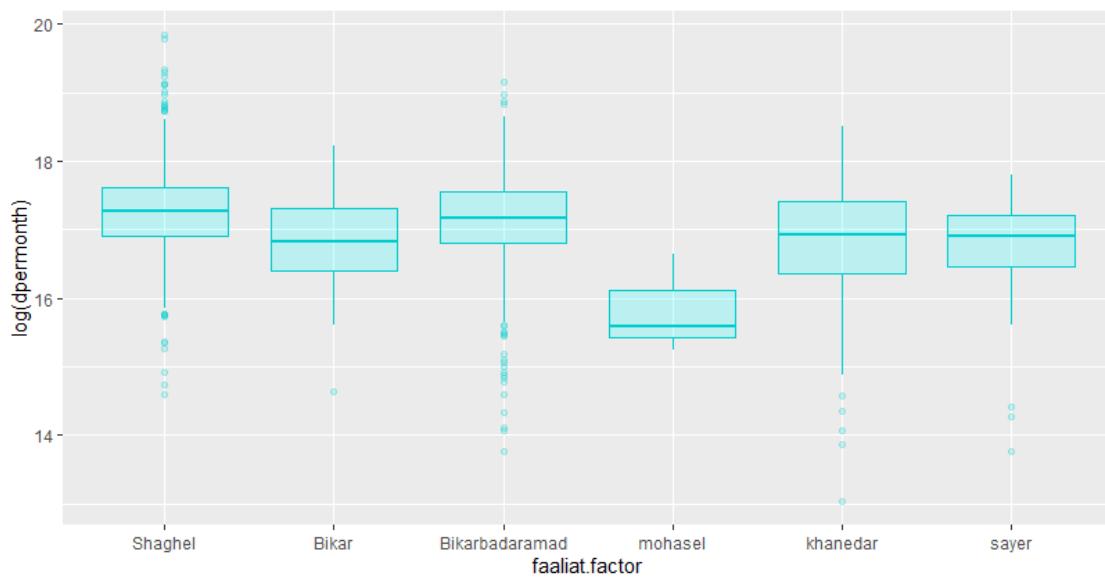


10) مصالح ساختمانی و درآمد



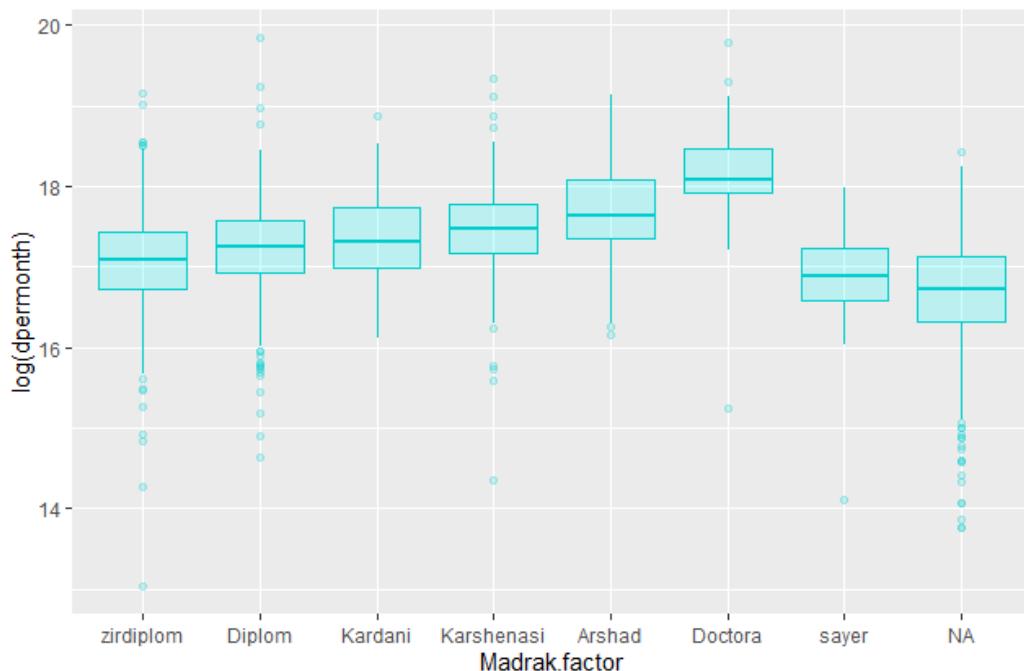
بنظر میرسد افرادی که نوع مصالح خشتی و گلی دارند درآمد پایین تری دارند.

### 11) فعالیت و لگاریتم درآمد



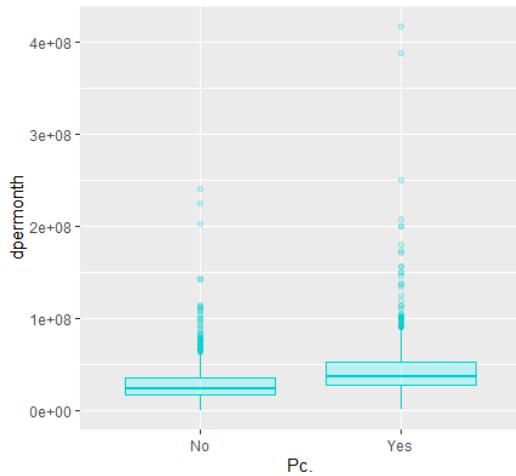
همانطور که انتظار داشتیم درآمد افراد شاغل از سایر گروهها بیشتر است.

### 12) مدرک تحصیلی و لگاریتم درآمد



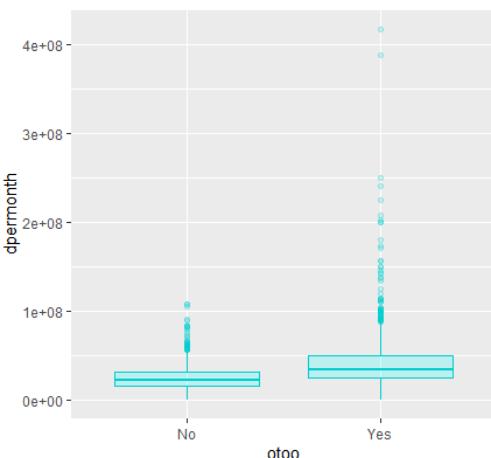
درآمد با افزایش سطح تحصیلات افزایش پیدا می‌کند

### 13) کامپیوتر و درآمد



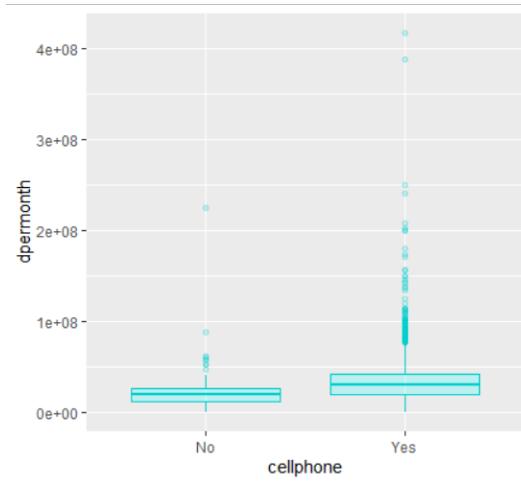
افرادی که صاحب کامپیوتر هستند درآمد بیشتری دارند.

### 14) ماشین و درآمد



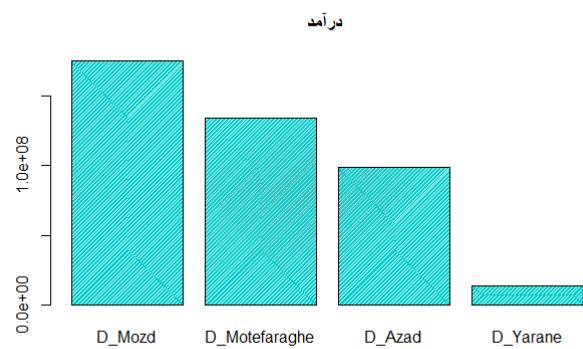
افرادی که صاحب ماشین هستند درآمد بیشتری دارند.

### 15) درآمد و تلفن همراه



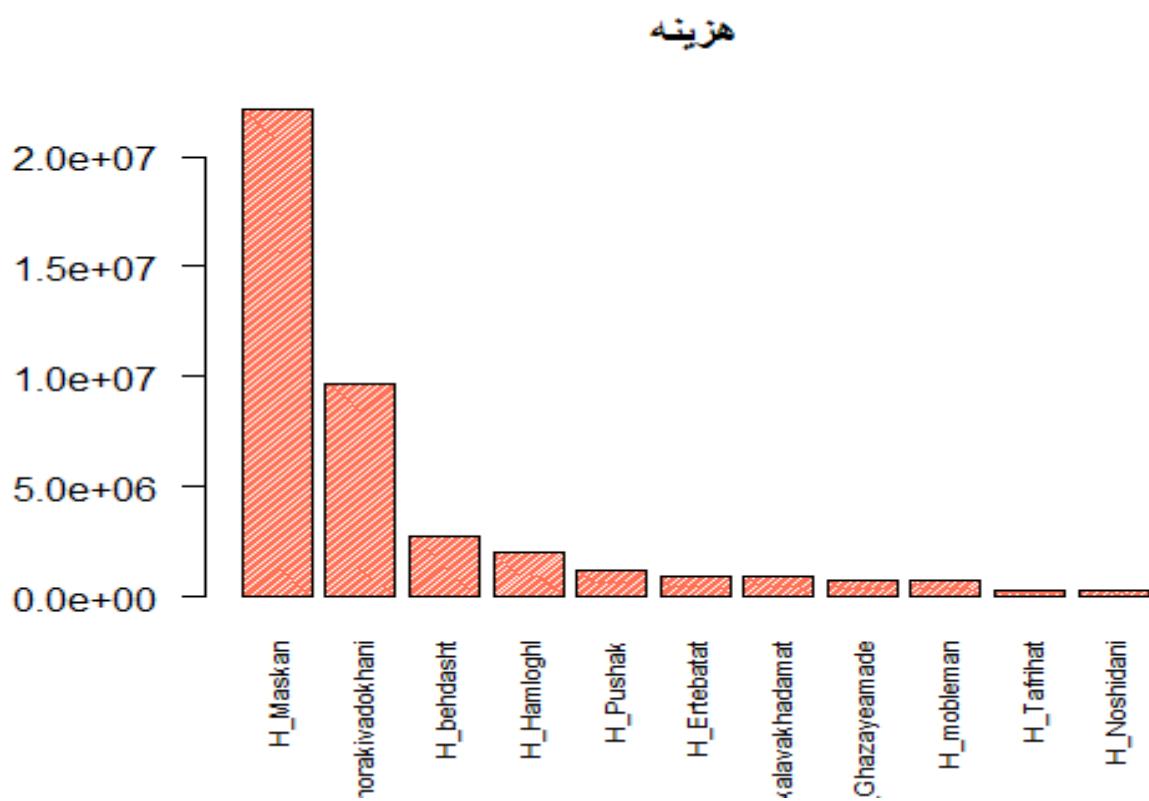
افرادی که صاحب تلفن همراه هستند درآمد بیشتری دارند.

12) مقایسه میانگین انواع درآمد



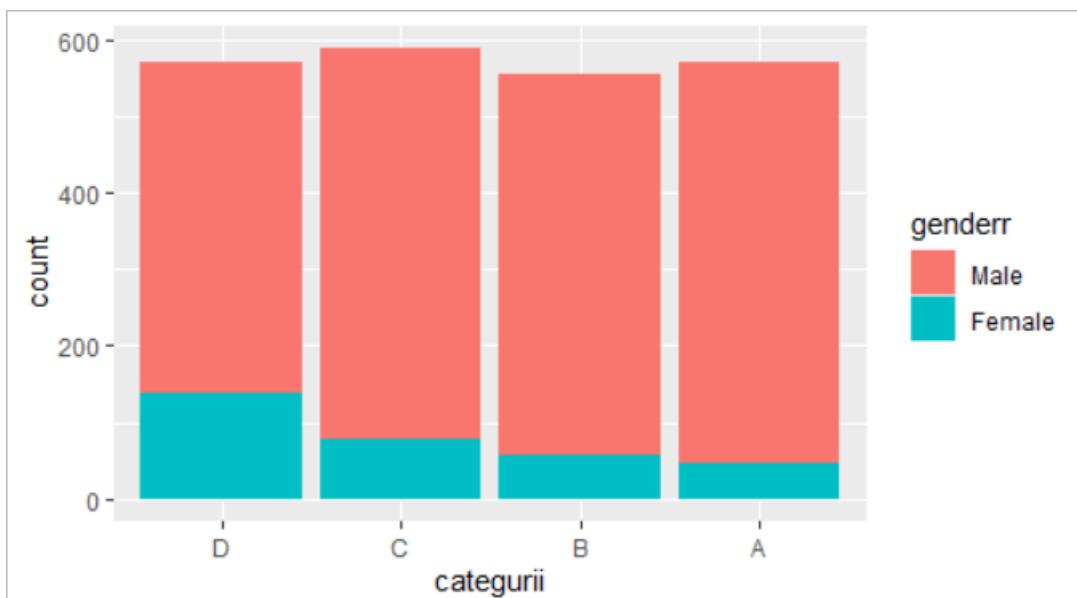
بیشترین بخش درآمد از بخش درآمد مزد است.

13) مقایسه میانگین انواع هزینه

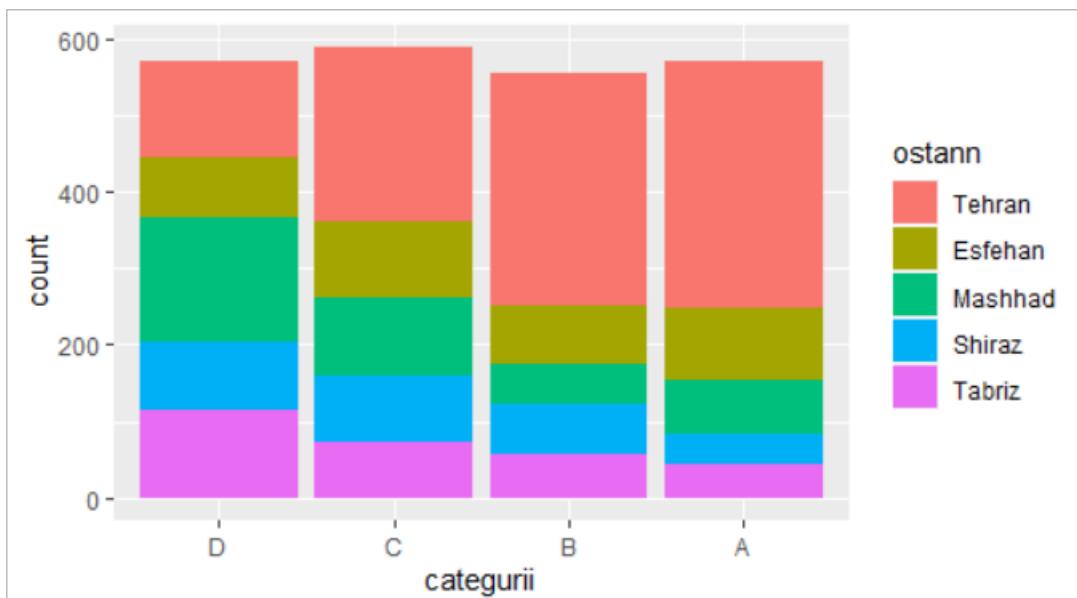


همانطور که در نمودار انواع هزینه مشاهده می‌کنید بیشترین هزینه مربوط به هزینه مسکن و خوراکی است.

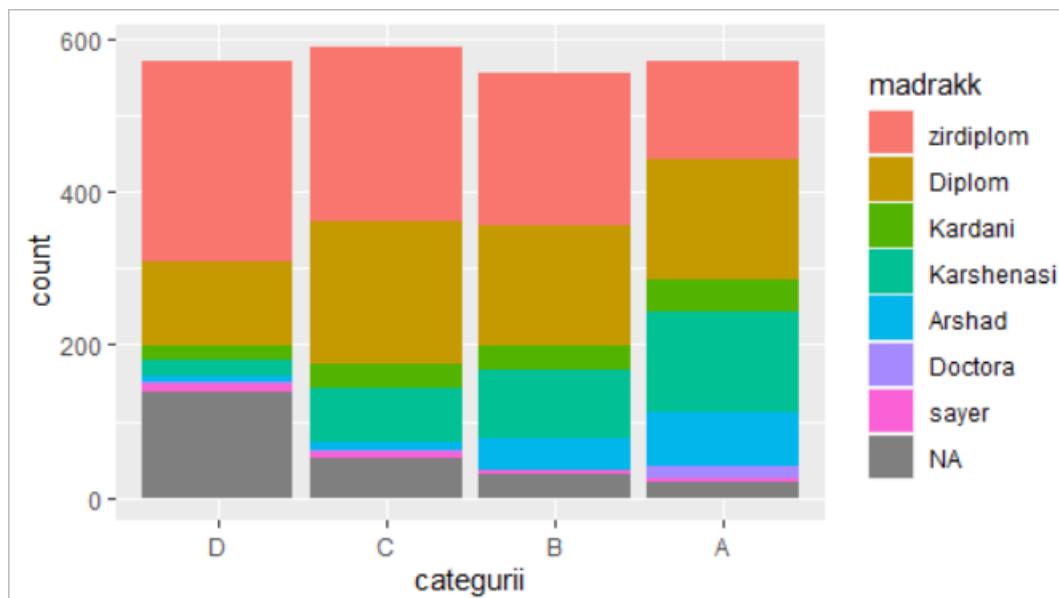
14) نمودار فراوانی رده‌های درآمدی به تفکیک جنسیت



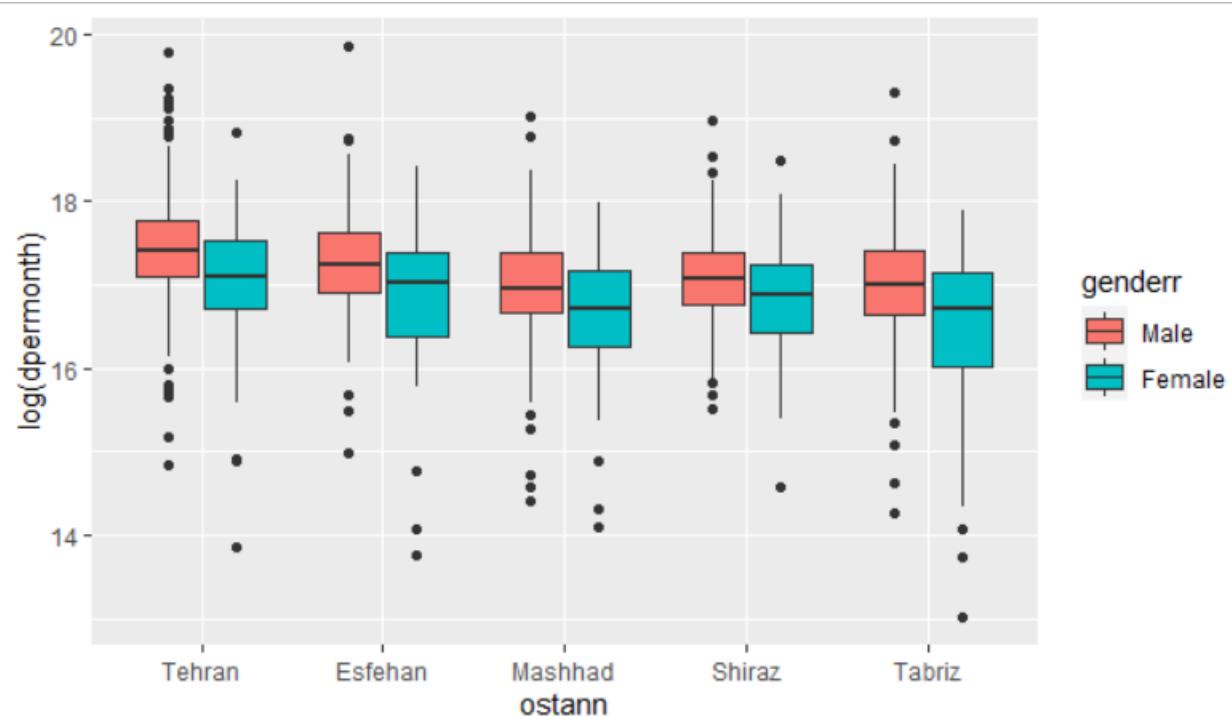
15) نمودار فراوانی رده‌های درآمدی به تفکیک استان



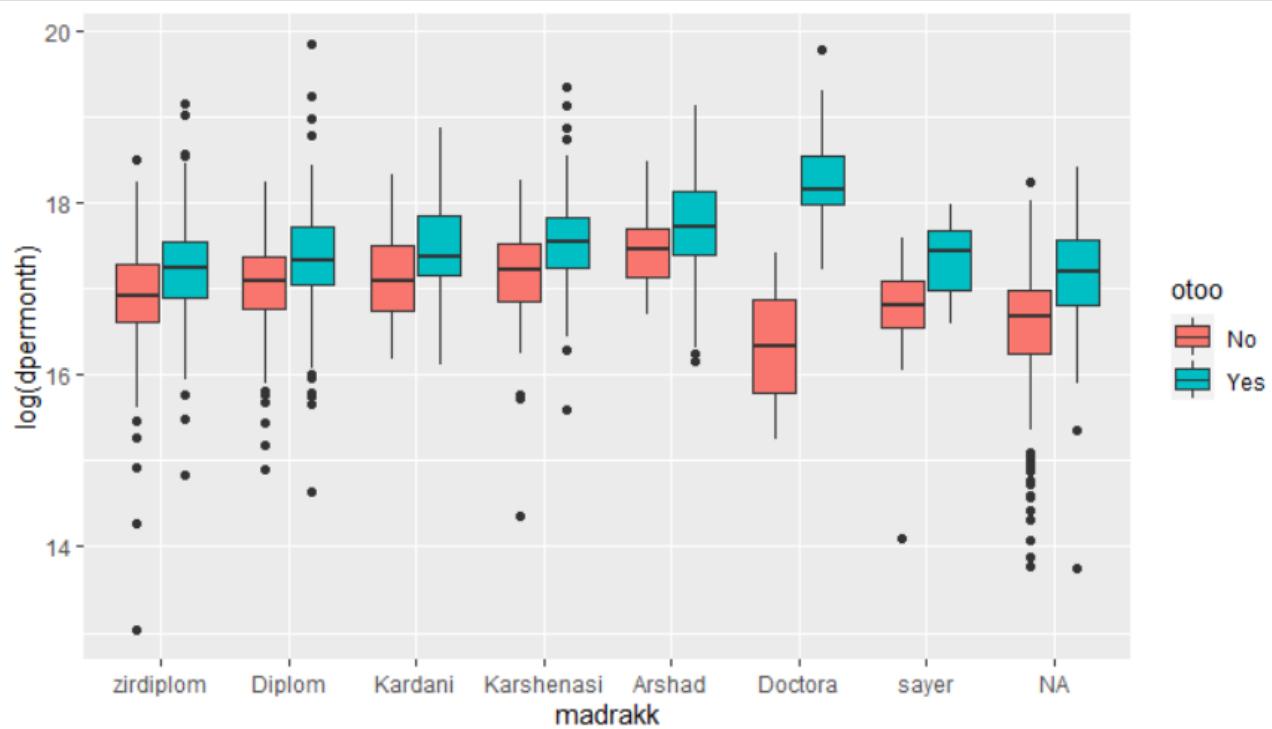
15) نمودار فراوانی رده‌های درآمدی به تفکیک مدرک تحصیلی



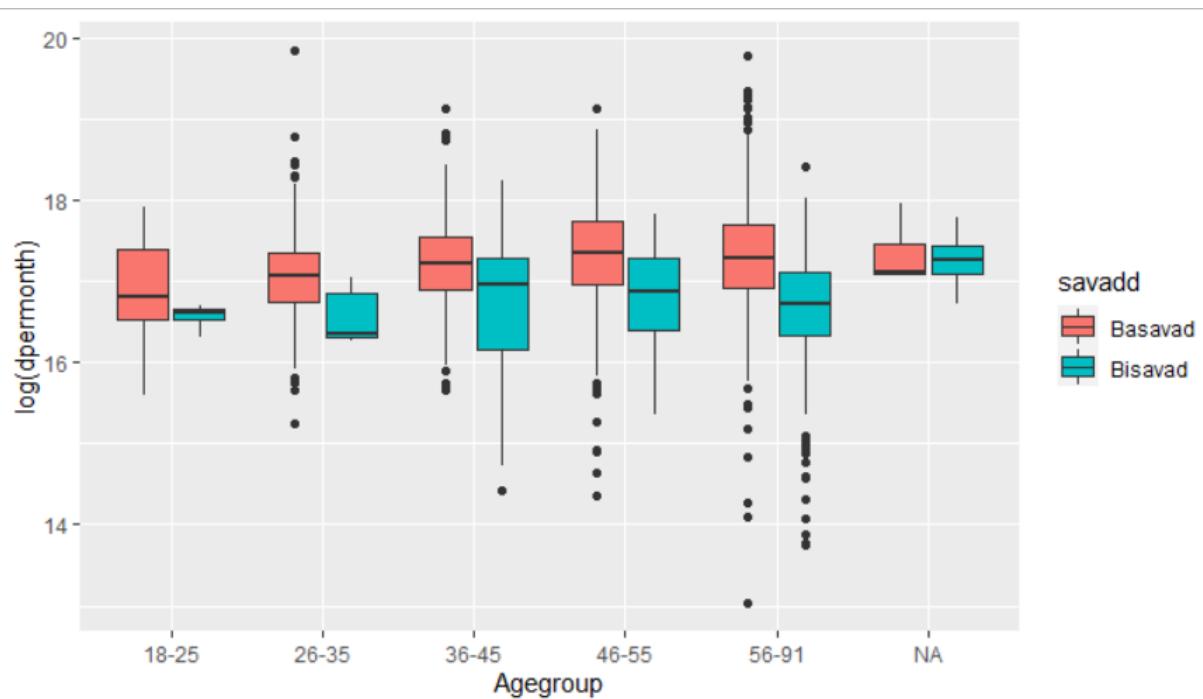
16) نمودار لگاریتم درآمد در مقابل استان به تفکیک جنسیت



17) نمودار لگاریتم درآمد مقابل مدرک تحصیلی به تفکیک داشتن یا نداشتن اتومبیل

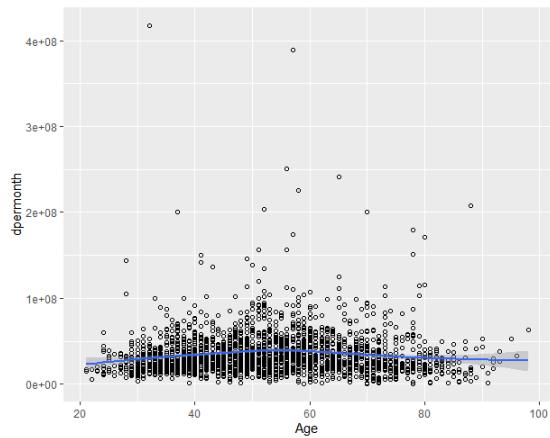


18) نمودار لگاریتم درآمد مقابل گروه سنی به تفکیک سواد



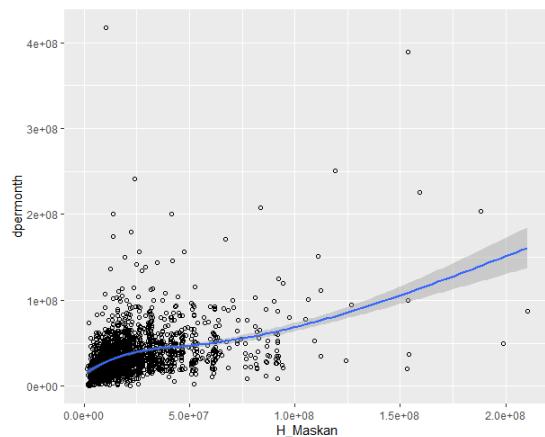
#### د) نمودار پراکنش

1) سن و درآمد



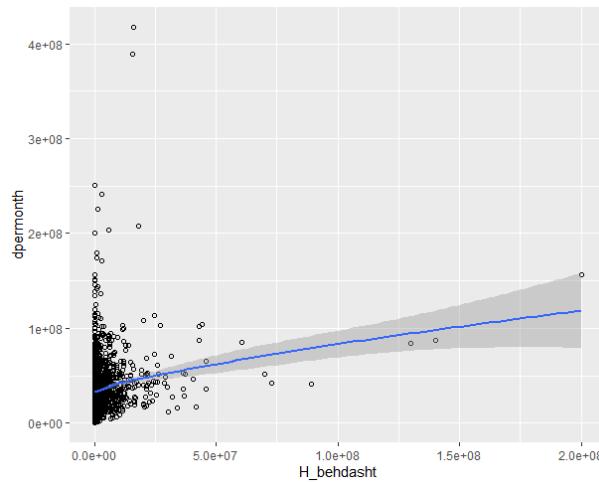
در این نمودار متوجه می‌شویم که همبستگی بین سن و درآمد ضعیف است

2) هزینه مسکن و درآمد



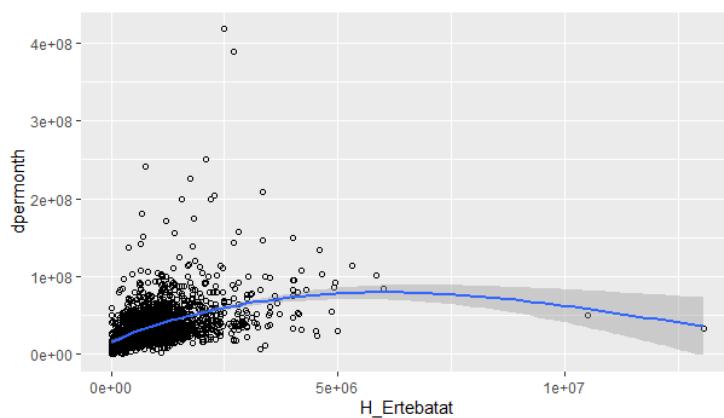
هرچه هزینه مسکن بیشتر باشد درآمد فرد نیز بیشتر است.

### (3) هزینه بهداشت و درآمد



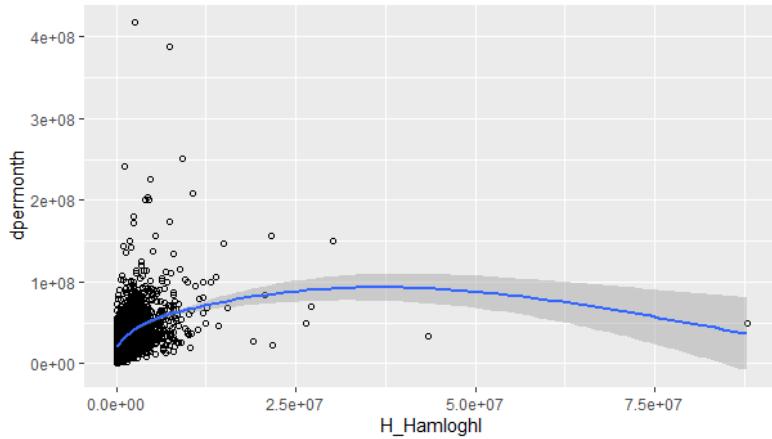
هرچه هزینه بهداشت بیشتر باشد درآمد فرد نیز بیشتر است.

### (4) هزینه ارتباطات و درآمد



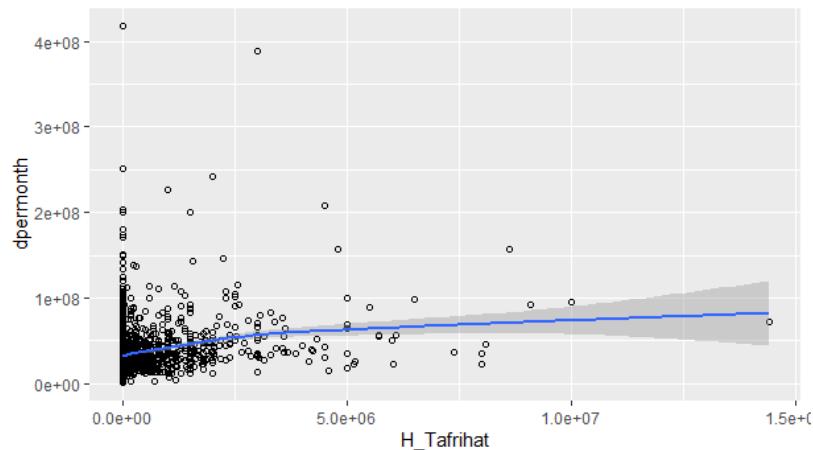
هرچه هزینه ارتباطات بیشتر باشد درآمد فرد نیز بیشتر است البته تعدادی داده دورافتاده وجود دارد که با حذف آن همبستگی بیشتری را شاهد خواهیم بود

##### 5) هزینه حمل و نقل و درآمد



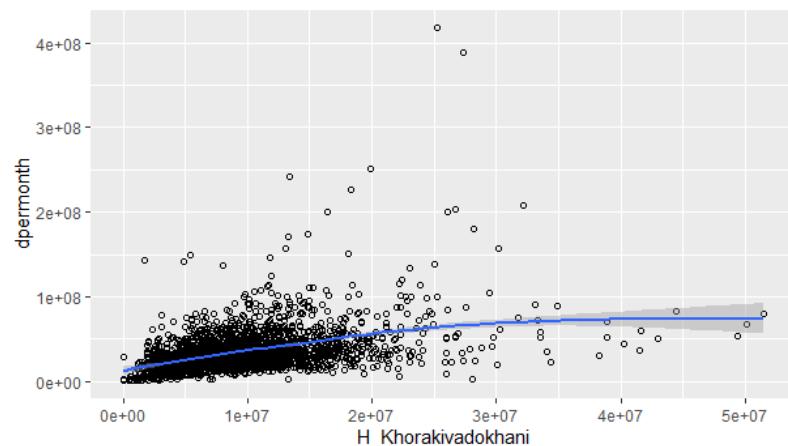
هرچه هزینه حمل و نقل بیشتر باشد درآمد فرد نیز بیشتر است البته تعدادی داده دورافتاده وجود دارد که با حذف آن همبستگی بیشتری را شاهد خواهیم بود

##### 6) هزینه تفریحات و درآمد



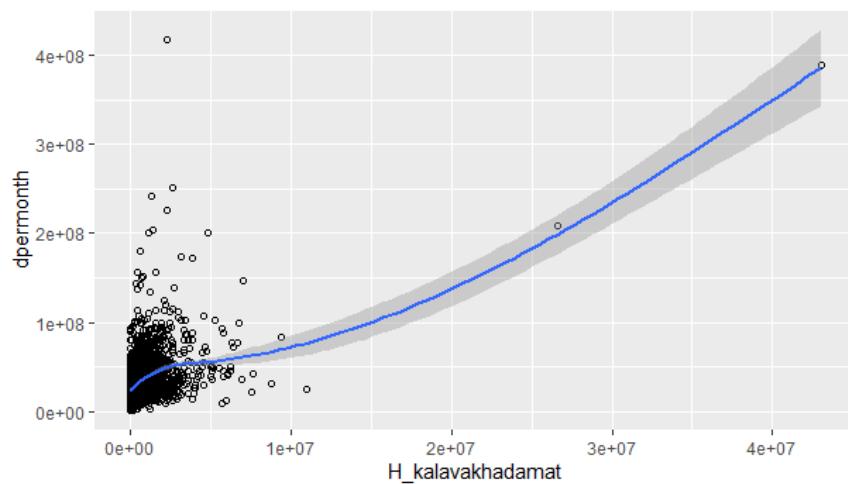
هرچه هزینه تفریحات بیشتر باشد درآمد فرد نیز بیشتر

7) هزینه خوراکی و دخانیات با درآمد



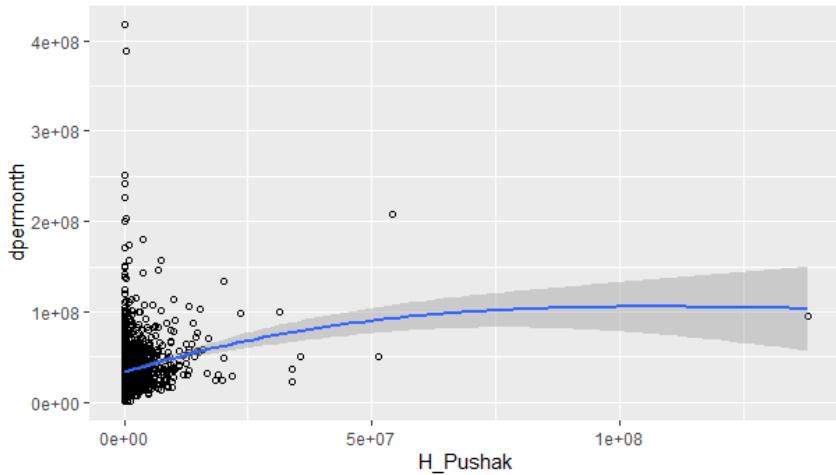
هرچه هزینه خوراکی و دخانیات بیشتر باشد درآمد فرد نیز بیشتر

8) هزینه کالا و خدمات و درآمد



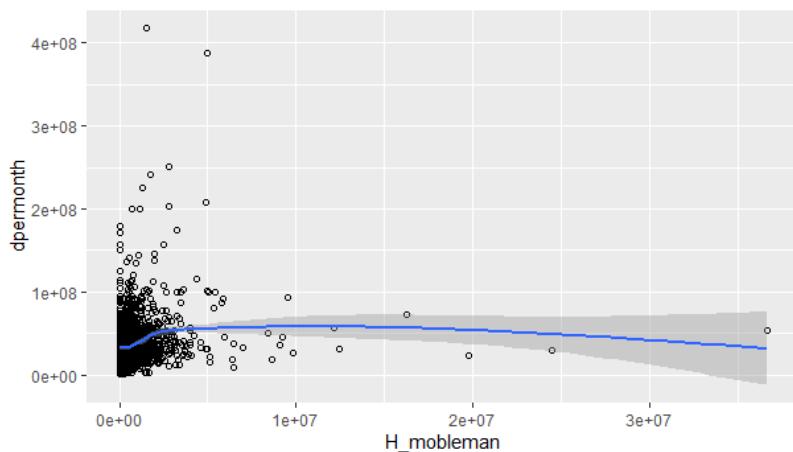
هرچه هزینه کالا و خدمات بیشتر باشد درآمد فرد نیز بیشتر

9) هزینه پوشак و درآمد



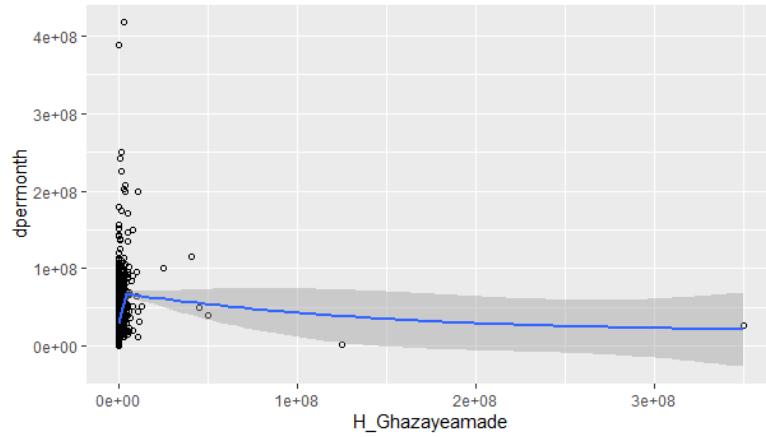
هرچه هزینه پوشак بیشتر باشد درآمد فرد نیز بیشتر

10) هزینه مبلمان و درآمد



در نمودار فوق همبستگی ضعیفی بین درآمد و هزینه مبلمان وجود دارد.

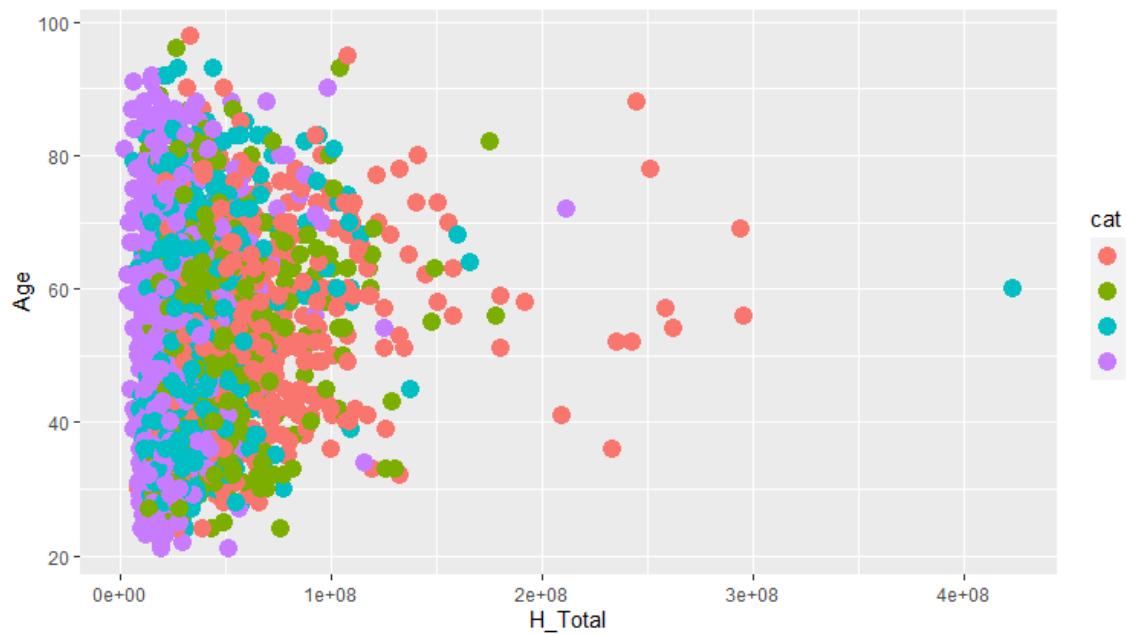
11) هزینه غذای آمده و درآمد



بنظر می‌آید ارتباطی بین هزینه غذای آمده و درآمد فرد وجود ندارد

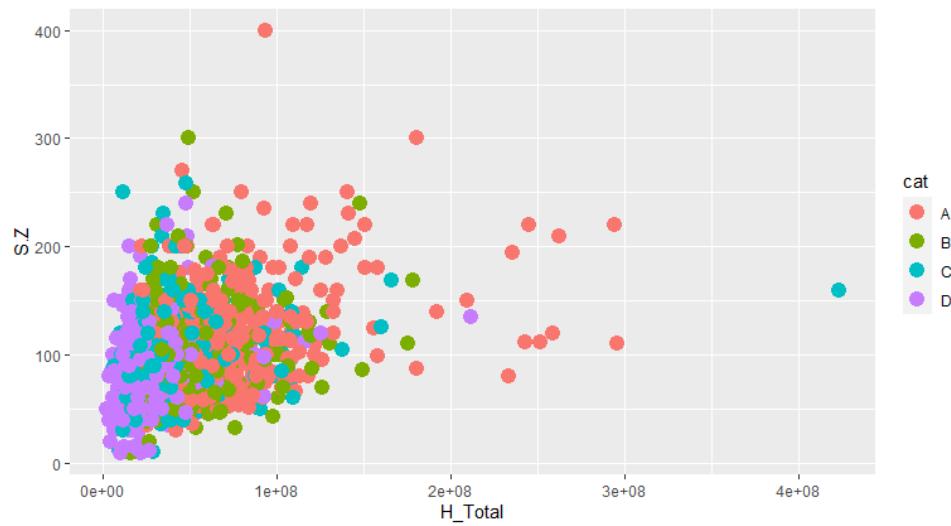
۵) نمودارهای سه بعدی

1) نمودار سن و هزینه کل و گروه بندی درآمدی



بهوضوح در نمودار مشخص هستکه گروه های پردرامد تر هزینه مسکن و سن بیشتری دارند

2) نمودار هزینه کل و سطح زیربنای ساختمان

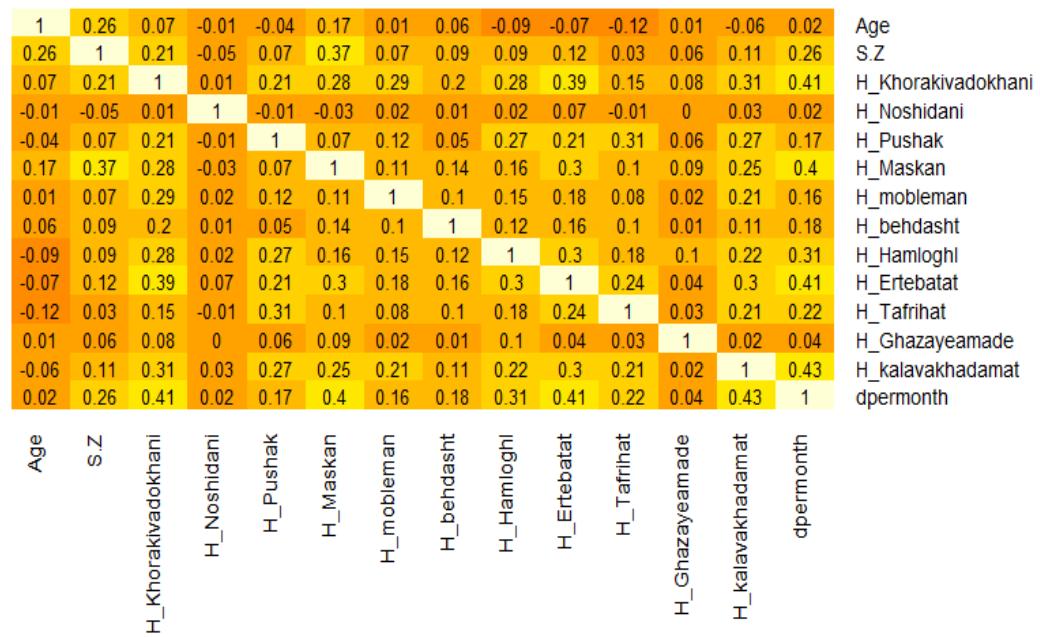


نمودار بالا به زیبایی نشان می دهد که کسانی که درآمد کم و سطح زیر بنای کمتری دارند در گروه D که با بنش نمایش داده شده اند قرار دارند

### (3) نمودار حرارتی

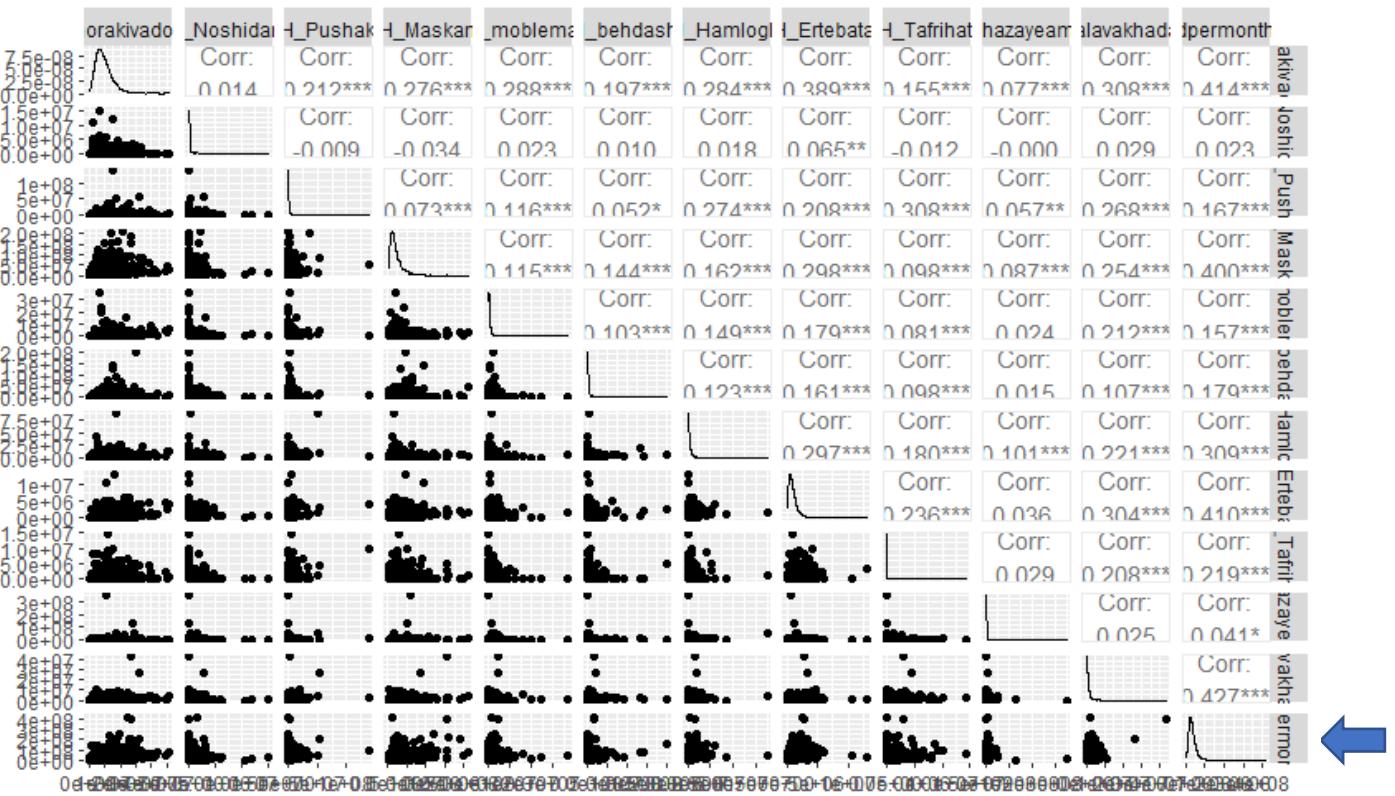
الف ) متغیرهای پیوسته

#### نمودار حرارتی



بر اساس نمودار بالا درآمد ماهانه با هزینه کالا و خدمات، ارتباطات، خوراکی و دخانیات همبستگی مثبت قوی تری دارد

4) ماتریس نمودار پراکنش هزینه و درآمد ماهانه



همانطور که در نمودار بالا ملاحظه می‌منید ردیف اخر که با فلش مشخص شده است مربوط به متغیر پاسخ یعنی درآمد ماهانه است و با متغیرهای هزینه مسکن، هزینه کالا و خدمات، هزینه خوراک ارتباط بظاهر خطی دارد.

## نتایج بخش تصویری سازی

برخی متغیرهایی که دارای تعداد زیادی مقادیر گمشده هستند که آنها را نیز از مجموعه داده کلی حذف میکنیم تا با متغیرهای کمتری سروکار داشته باشیم و تحلیل درست‌تری انجام دهیم، اما متغیرهای کاندید حذف شدن و غیرموثر بر درآمد :

متغیرهای مصالح ساختمانی، تلویزیون، دستگاه DVD، اجاق گاز، پنکه، فاضلاب حذف می‌شوند و متغیرهای هزینه به علت اینکه پاسخگو جوابی درست نمی‌تواند بدهد را به جز متغیرهای هزینه کالا و خدمات، ارتباطات، مسکن و خوارکی بقیه متغیرهای هزینه را حذف می‌کنیم.

فراخوانی و افزار داده‌ها

```
library(readxl)
data<-read_excel("C:/Users/Administrator/Desktop/data.xlsx")
data[,c(1,3,5,6,7,8,9,12:36,42)] <- lapply(data[,c(1,3,5,6,7,8,9,12:36,42)], factor)
new.data=data[,-c(37:41)]
View(new.data)
attach(data)

set.seed(1234)
train.row=sample(rownames(new.data),dim(new.data)[1]*0.5)
valid.row=sample(setdiff(rownames(new.data),train.row),dim(new.data)[1]*0.35)
test.row=setdiff(rownames(new.data),union(train.row,valid.row))
train.data=new.data[train.row,]
valid.data=new.data[valid.row,]
test.data=new.data[test.row,]
```

## درخت رده‌بندی

از انجایی که هدف این پروژه رده بندی است یکی دیگر از مدل‌هایی که می‌تواند به ما کمک کند، مدل درخت است که بدلیل عملکرد آن یکی از شفافترین و محبوب‌ترین روشها برای تفسیر متغیرها شناخته می‌شود. در ادامه بیشتر با این روش آشنا می‌شویم

پس از افزار داده‌ها به داده‌های آموزشی و اعتبارسنجی مدل را ابتدا روی داده‌های آموزشی پیدا‌سازی می‌کنیم و بعد در بخش اعتبارسنجی به ارزیابی آن می‌پردازیم

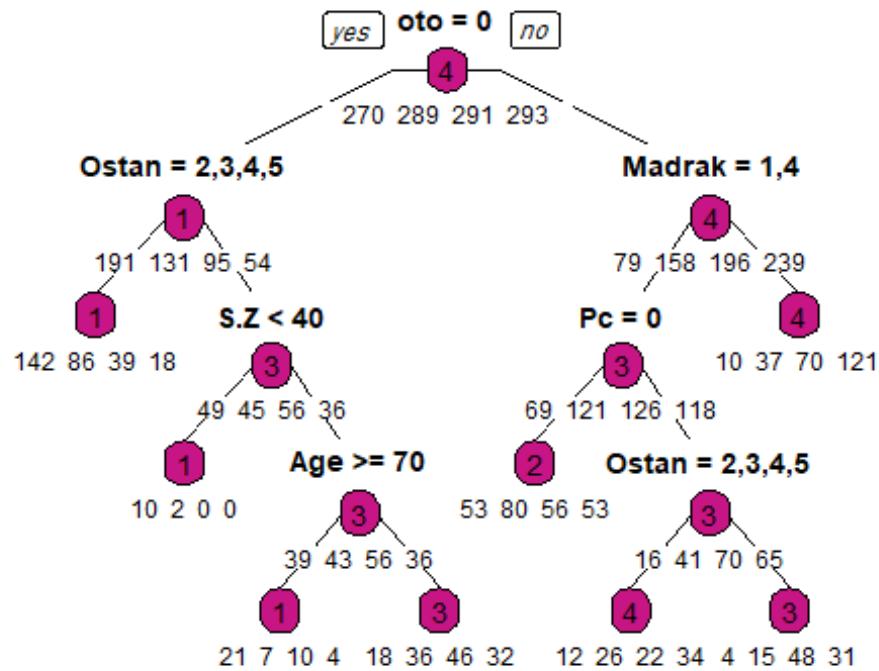
کتابخانه‌های مورد نیاز

#TREE

```
library(rpart)
library(rpart.plot)
library(caret)
library(lattice)
library(ggplot2)
library(e1071)
```

سپس در این مرحله مدل درخت را برازش داده و درخت رده بندی را رسم می‌کنیم

```
class.tree <- rpart( categori~, data = train.data, method = "class", model=TRUE )
prp(class.tree, type = 1, extra = 1, under = TRUE, split.font = 2, varlen = -10,
cex=.9,box.palette=c("mediumvioletred"))
```



سپس به بررسی دقت مدل میپردازیم

## ماتریس درهم ریختگی داده های آموزشی

```

class.tree.t <- predict(class.tree, train.data, type = "class")
confusionMatrix(as.factor(class.tree.t), as.factor(train.data$categori))

## Confusion Matrix and Statistics
##
##           Reference
## Prediction  1   2   3   4
##       1 173  95  49  22
##       2  53  80  56  53
##       3  22  51  94  63
##       4  22  63  92 155
##
## Overall Statistics
##
##           Accuracy : 0.4392
## 95% CI : (0.41017, 0.46853)
## No Information Rate : 0.25634
## P-Value [Acc > NIR] : < 2.22e-16
##
##           Kappa : 0.25298
## 

```

```

## McNemar's Test P-Value : 6.9084e-05
##
## Statistics by Class:
##
##                               Class: 1 Class: 2 Class: 3 Class: 4
## Sensitivity             0.64074  0.276817  0.32302   0.52901
## Specificity            0.80985  0.810304  0.84038   0.79176

```

مقدار درستی برابر 0.44 است و مدل در تشخیص رده 3 و 2 عملکرد ضعیفتری دارد.

## ماتریس درهم ریختگی برای داده‌های اعتبار سنجی

```

class.tree.v <- predict(class.tree,valid.data,type = "class")
confusionMatrix(as.factor(class.tree.v),as.factor(valid.data$categori))

## Confusion Matrix and Statistics

## Reference

## Prediction 1 2 3 4
##          1 126 67 34 17
##          2 54 45 41 30
##          3 17 45 53 49
##          4 16 49 55 102

##

## Overall Statistics

##

## Accuracy : 0.4075
## 95% CI : (0.37322, 0.44246)
## No Information Rate : 0.26625
## P-Value [Acc > NIR] : < 2e-16

##

## Kappa : 0.20842
## 

## McNemar's Test P-Value : 0.057747

```

```

## 
## Statistics by Class:
## 
##          Class: 1 Class: 2 Class: 3 Class: 4
## Sensitivity      0.59155  0.21845  0.28962  0.51515
## Specificity      0.79898  0.78956  0.82010  0.80066

```

مقدار درستی در داده های اعتبار سنجی 0.41 است که کمتر از داده های آموزشی است.

## درخت هرس شده

```

set.seed(1234)
deeper.tree<- rpart(categuri ~ ., data = train.data, method = "class", cp= .00001
, minsplit =5)
options(digits=8)
printcp(deeper.tree)

## 
## Classification tree:
## rpart(formula = categuri ~ ., data = train.data, method = "class",
##       cp = 1e-05, minsplit = 5)
##
## Variables actually used in tree construction:
## [1] Age         charkh.kh cooler      do          Faaliat    freezer   Gender     ha
## mam        hararat.m internet   m.lebas
## [12] Madrak     microfer   motor      N.S        Ostan      oto        package  Pc
## radio      S.Z        T.M.S
## [23] T.O        Tedad      yakhchal  zabt
##
## Root node error: 850/1143 = 0.743657
##
## n= 1143
##
##           CP nsplit rel.error   xerror      xstd
## 1  0.161176471      0  1.000000 1.065882 0.0161249
## 2  0.018823529      1  0.838824 0.838824 0.0192680
## 3  0.014117647      3  0.801176 0.855294 0.0191369
## 4  0.010980392      4  0.787059 0.844706 0.0192227
## 5  0.009411765      7  0.754118 0.817647 0.0194173
## 6  0.008235294      9  0.735294 0.812941 0.0194476
## 7  0.007058824     10  0.727059 0.810588 0.0194624
## 8  0.005882353     13  0.705882 0.834118 0.0193031
## 9  0.004705882     15  0.694118 0.825882 0.0193618

```

```

## 10 0.003529412      22 0.661176 0.832941 0.0193116
## 11 0.003137255      26 0.647059 0.835294 0.0192944
## 12 0.002941176      43 0.582353 0.844706 0.0192227
## 13 0.002745098      52 0.555294 0.840000 0.0192591
## 14 0.002352941      55 0.547059 0.835294 0.0192944
## 15 0.001960784      98 0.442353 0.840000 0.0192591
## 16 0.001764706     125 0.388235 0.848235 0.0191947
## 17 0.001568627     144 0.351765 0.852941 0.0191565
## 18 0.001470588     152 0.338824 0.850588 0.0191757
## 19 0.001176471     158 0.329412 0.850588 0.0191757
## 20 0.000784314     228 0.244706 0.872941 0.0189816
## 21 0.000588235     238 0.236471 0.872941 0.0189816
## 22 0.000294118     246 0.231765 0.877647 0.0189375
## 23 0.000010000     250 0.230588 0.877647 0.0189375

which.min(deeper.tree$cptable[, "xerror"])

## 7
## 7

```

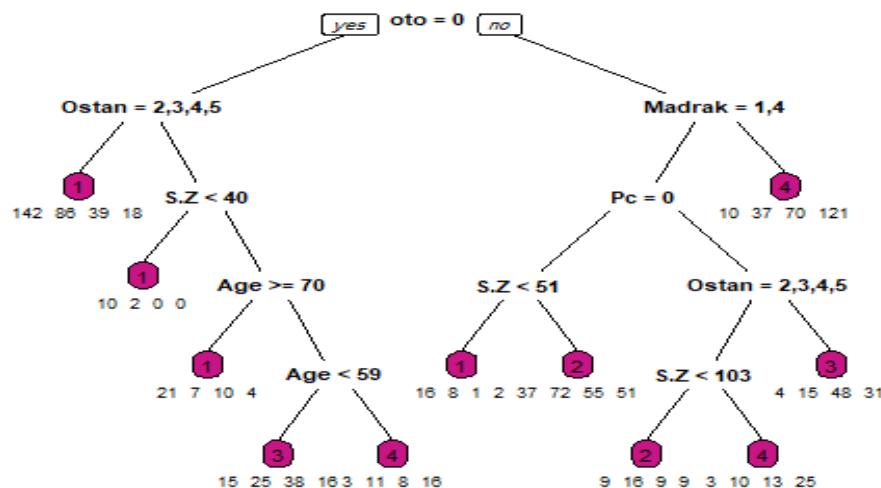
پس رديف 11 که مقدار cp ان برابر **0.007058824** است و تعداد 10 جداسازی دارد

دستورات ايجاد درخت هرس شده با cp بهينه شده در داده های آموزشی

```

cv.tree<-rpart(categuri~,data = train.data ,method = "class",
                 cp =0.007058824, minsplit =5,xval =5, model=T)
prp(cv.tree,type =0,extra = 1,under = T,split.font = 2,
    varlen = -10,cex=.7,box.palette=c("mediumvioletred"))

```



## ماتریس درهم ریختگی داده‌های آموزشی درخت هرس شده

```
cv.tree.p <- predict(cv.tree,train.data,type = "class")
confusionMatrix(as.factor(cv.tree.p),as.factor(train.data$categori))

## Confusion Matrix and Statistics
##             Reference
## Prediction 1 2 3 4
##       1 189 103 50 24
##       2 46 88 64 60
##       3 19 40 86 47
##       4 16 58 91 162
##
## Overall Statistics
##
##                 Accuracy : 0.45932
##                 95% CI : (0.43013, 0.48872)
## No Information Rate : 0.25634
## P-Value [Acc > NIR] : < 2.22e-16
## Kappa : 0.28023
## McNemar's Test P-Value : 1.8838e-10
## Statistics by Class:
##                 Class: 1 Class: 2 Class: 3 Class: 4
## Sensitivity      0.70000 0.30450 0.295533 0.55290
## Specificity      0.79725 0.80094 0.875587 0.80588
```

مقدار درستی در داده‌های آموزشی مدل درخت هرس شده ۰.۴۶ است که نسبت به درخت رده‌بندی اولیه افزایش داشته است

## ماتریس درهم ریختگی داده‌های اعتبارسنجی درخت هرس شده

```
cv.tree.p.v <- predict(cv.tree,valid.data,type = "class")
confusionMatrix(as.factor(cv.tree.p.v),as.factor(valid.data$categori))

## Confusion Matrix and Statistics
##
##             Reference
## Prediction 1 2 3 4
##       1 132 73 38 18
##       2 53 50 44 40
##       3 9 39 44 46
##       4 19 44 57 94
## Overall Statistics
```

```

##          Accuracy : 0.4
##          95% CI : (0.36585, 0.43489)
##          No Information Rate : 0.26625
##          P-Value [Acc > NIR] : < 2.22e-16
##          Kappa : 0.19709
##
## Statistics by Class:
##          Class: 1 Class: 2 Class: 3 Class: 4
## Sensitivity     0.61972  0.24272  0.24044  0.47475
## Specificity    0.78024  0.76936  0.84765  0.80066

```

مقدار درستی در داده‌های اعتبار سنجی برابر ۰.۴ است که نسبت به آموزشی کمتر است و در اینجا بیش برآذش داریم.

## مدل رگرسیون ترتیبی

همانطور که در فصل قبل اشاره شد رگرسیون یکی از روش‌های راهنماییده در داده‌کاوی است در این مسئله به علت اینکه متغیر پاسخ ما مشخص کننده‌ی سطوح درآمدی از کم درآمد تا پر درآمد ۴ رده است و به اصطلاح رسته‌ای ترتیبی است پس از رگرسیون ترتیبی استفاده می‌کنیم. در این قسمت مدل رگرسیونی را برآذش داده بر روی افرادهای داده از جمله آموزشی و اعتبار سنجی برآذش می‌دهیم.

```

library(ordinal)
model=polr(categuri~ Ostan+ Tedad+ Gender+ Age+ Savad+ Faaliat+ T.M.S+
T.O+ S.Z+ N.S +oto+ Pc+ mobile+ freezer+ yakhchal+ yakhchal.f+m.lebas+
charkh.kh+ m.zarf+ microfer+ internet+ hararat.m +package+H_Khorakivadolhani
+H_Maskan+H_Ertebatat+H_kalavakhadamat
, data=train.data ,na.action=na.omit,Hess=TRUE,method="logistic")

```

## ماتریس برهم ریختگی بر روی داده‌های آموزشی

مدل قسمت قبل را بر روی داده‌های آموزشی برآذش می‌دهیم.

```

pred<-predict(model,train.data,type="class")
confusionMatrix(pred,train.data$categuri)

```

```

# Confusion Matrix and Statistics
##             Reference
## Prediction 1 2 3 4
##           1 165 65 18 4
##           2 75 127 92 39
##           3 23 61 89 89
##           4 7 36 92 161
## Overall Statistics
##                 Accuracy : 0.47419
##                 95% CI : (0.44449, 0.50361)
## No Information Rate : 0.25634
## P-Value [Acc > NIR] : < 2e-16
##
##                 Kappa : 0.29847
##
## McNemar's Test P-Value : 0.19779
## Statistics by Class:
##                  Class: 1 Class: 2 Class: 3 Class: 4
## Sensitivity       0.61111 0.43945 0.305842 0.54949
## Specificity       0.90034 0.75878 0.796948 0.84118

```

همانطور که از نتایج مشخص است مقدار درستی مدل بر روی داده های آموزشی 0.474 است و مدل تشخیص رده های 2 و 3 همچنان ضعیف عمل می کند زیرا مقدار حساسیت در این دو رده کمتر از رده 1 و 4 است.

## ماتریس برهم ریختگی بر روی داده های اعتبارسنجی

```

pred<-predict(model,valid.data,type="class")
confusionMatrix(pred,valid.data$categori)

# Confusion Matrix and Statistics
##             Reference
## Prediction 1 2 3 4
##           1 127 45 21 3
##           2 57 80 46 33
##           3 26 61 69 58
##           4 3 20 47 104
##
## Overall Statistics
##
##                 Accuracy : 0.475
##                 95% CI : (0.43992, 0.51027)
## No Information Rate : 0.26625

```

```

##      P-Value [Acc > NIR] : < 2e-16
##
##          Kappa : 0.30021
##
##  Mcnemar's Test P-Value : 0.21106
##
## Statistics by Class:
##
##                               Class: 1 Class: 2 Class: 3 Class: 4
## Sensitivity           0.59624  0.38835  0.37705  0.52525
## Specificity          0.88245  0.77104  0.76499  0.88372

```

مقدار درستی در داده های اعتبار سنجی تقریبا با داده های آموزشی یکسان است

## SVM روش

همانطور که در فصل پیش توضیح داده شد SVM یکی از روش های یادگیری ماشین است و یادگیری بانظارت است همچنین الگوریتم SVM، جزو الگوریتم های تشخیص الگو دسته بندی می شود. از الگوریتم SVM در هر جایی که نیاز به تشخیص الگو یا دسته بندی اشیا در کlassen های خاص باشد می توان استفاده کرد. در اینجا به کمک این روش می خواهیم طبقه اقتصادی اجتماعی را به کمک آن تعیین کنیم.

```

library(e1071)
library(caret)
attach(data.df)
model=svm(categori~.,data = train,type="C-classification")
summary(model)
Call:
svm(formula = categori ~ ., data = train.svm, type = "C-classification")

```

Parameters:

SVM-Type: C-classification

SVM-Kernel: radial

cost: 1

Number of Support Vectors: 1279

( 338 315 369 257 )

## ماتریس برهم ریختگی بر روی داده های آموزشی

confusionMatrix(pred,cate)

Confusion Matrix and Statistics

Reference

Prediction 1 2 3 4

18 43 104 267 1

55 102 154 55 2

32 91 27 10 3

253 79 65 17 4

Overall Statistics

Accuracy : 0.5504

.95

CI(0.5841 ,0.5308) :

No Information Rate : 0.2609

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.4072

McNemar's Test P-Value : < 2.2e-16

مقدار درستی در داده های آموزشی 0.55 است

Statistics by Class:

Class: 1 Class: 2 Class: 3 Class: 4

Sensitivity      0.7650 0.4400 0.28889 0.7067

Specificity      0.8387 0.7926 0.93472 0.8412

با توجه به مقدار حساسیت مدل میزان پیش‌بینی گروه 1 و 4 بیشتر از سایر گروه‌های درآمدی است

## ماتریس برهم ریختگی بر روی داده‌های اعتبارسنجی

```
pred <- predict(model.v, x1)
confusionMatrix(pred, categ)
```

Confusion Matrix and Statistics

Reference

Prediction	1	2	3	4
1	146	61	21	6
2	47	131	57	45
3	22	34	146	57
4	8	13	15	106

Overall Statistics

Accuracy : 0.5760

```
95% CI : (0.5454, 0.6104)
No Information Rate : 0.2612
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.4356

McNemar's Test P-Value : 4.505e-09

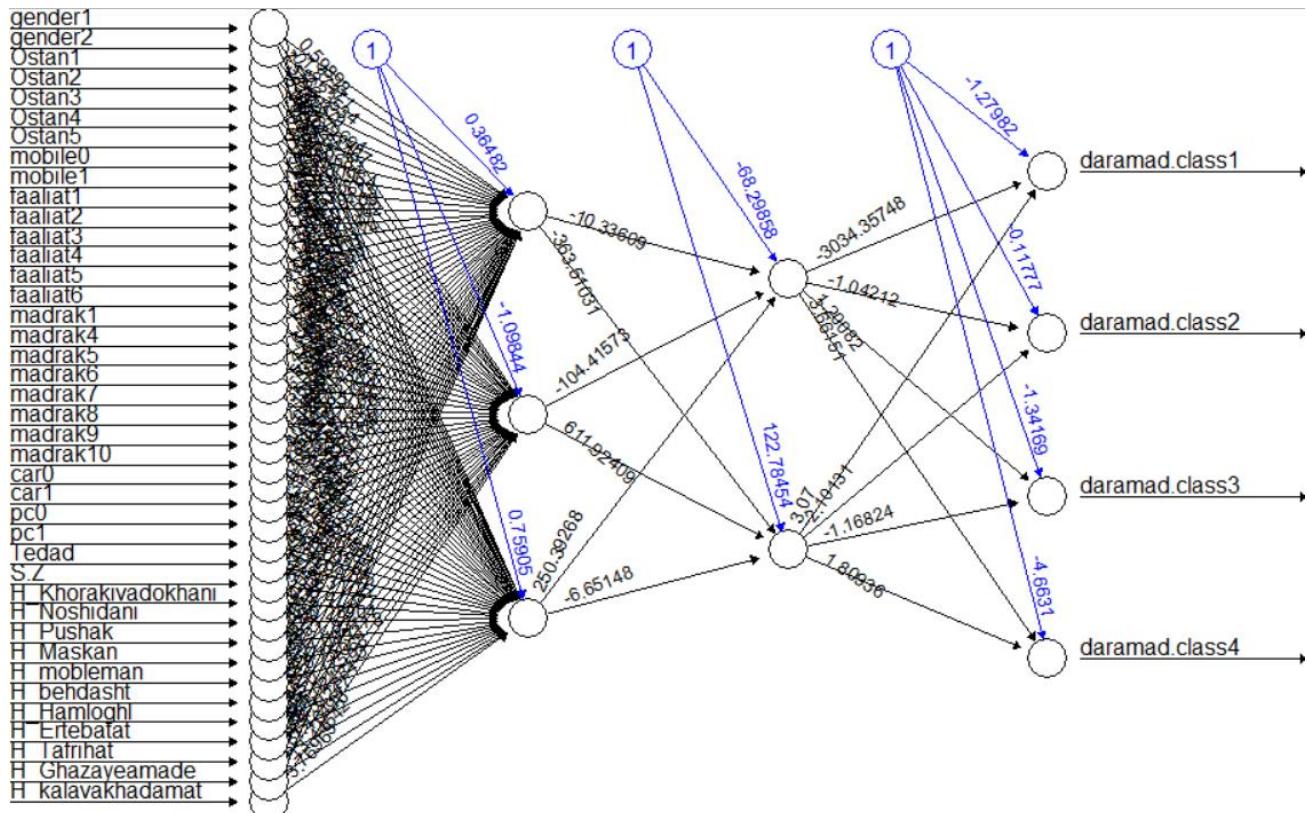
Statistics by Class:

          Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity      0.6547   0.5471   0.6119   0.4953
Specificity      0.8728   0.7796   0.8328   0.9486
```

میزان درستی در این مدل 0.5760 است که در مقایسه با داده‌های آموزشی مقداری افزایش یافته.

## شبکه عصبی

همانطور که در فصل قبل توضیح دادیم شبکه عصبی مدل هایی برای رده بندی و پیش گویی هستند که بر یک مدل فعالیت بیولوژی در مغز، مبتنی است که در آن سلول های عصبی در اتصال با یکدیگرند و از تجربه یاد می گیرند. در ادامه نمودار شبکه عصبی رسم شده است.



## مراحل اجرای شبکه عصبی

در ابتدا متغیرهای رسته ای که در تصویری سازی و بخش های قبل شناسایی شده است را به اصطلاح دامی می کنیم و متغیرهای پیوسته را نیز نرمال سازی می کنیم و بخش های آموزشی و اعتبار سنجی را مشخص می کنیم

الف) دامی کردن متغیرها

```
Gender=as.factor(Gender)
Ostan=as.factor(Ostan)
Faaliat=as.factor(Faaliat)
Madrak=as.factor(Madrak)
oto=as.factor(oto)
Pc=as.factor(Pc)
mobile=as.factor(mobile)
daramad.class=as.factor(categuri)
x1=class.ind(Gender)
x2=class.ind(Ostan)
x3=class.ind(mobile)
x4=class.ind(Faaliat)
x5=class.ind(Madrak)
x6=class.ind(oto)
x7=class.ind(Pc)
x8=class.ind(daramad.class)
dums=data.frame(x1,x2,x3,x4,x5,x6,x7,x8)

colnames(dums)=c(paste("gender",c(1,2),sep=""),
                 paste("Ostan",c(1,2,3,4,5),sep=""),
                 paste("mobile",c(0,1),sep=""),
                 paste("faaliat",c(1,2,3,4,5,6),sep=""),
                 paste("madrak",c(1,4,5,6,7,8,9,10),sep=""),
                 paste("car",c(0,1),sep=""),
                 paste("pc",c(0,1),sep=""),
                 paste("daramad.class",c(1,2,3,4),sep=""))
```

ب) نرمال سازی دادهها

```
norm.values <- preProcess(newdata[,c(2,9,40:42)], method=c("center",  
"scale"))  
newdata[,c(2,9,40:50)] <- predict(norm.values, newdata[,c(2,9,40:50)])  
data.nn=cbind(dums,newdata[,c(2,9,40:42)])
```

چون مقدار `set.seed` را برابر 1234 قرار دادیم پس نمونه ما برابر نمونه های مدل های قبلی می باشد

```
train.row=sample(rownames(data.nn),dim(data.nn)[1]*0.5)  
valid.row=sample(setdiff(rownames(new.data),train.row),dim(new.data)[1]*0.35)  
test.row=setdiff(rownames(data.nn),union(train.row,valid.row))  
train.data=new.data[train.row,]  
valid.data=new.data[valid.row,]  
test.data=new.data[test.row,]
```

ج) برازش مدل

```
library(nnet)  
library(neuralnet)  
nn=neuralnet(daramad.class1+daramad.class2+daramad.class3+daramad.class4~  
gender1 +gender2+ Ostan1+ Ostan2+ Ostan3+ Ostan4+ Ostan5 +mobile0+  
mobile1+ faaliat1+ faaliat2+ faaliat3+ faaliat4  
+faaliat5 +faaliat6 +madrak1 +madrak4 +madrak5  
+madrak6+ madrak7+ madrak8+ madrak9+ madrak10+ car0 +car1+  
pc0+ pc1+ Tedad+ S.Z+ H_Khorakivadokhani +H_Noshidani  
+H_Pushak+ H_Maskan+ H_mobleman+ H_behdasht+ H_Hamloghl  
+H_Ertebatat+ H_Tafrihat+ H_Ghazayeamade +H_kalavakhadamat  
,data = train.data, linear.output = F, hidden =c(3,2))
```

## برازش مدل شبکه عصبی و ماتریس برهم ریختگی بر روی داده‌های آموزشی

*train*

```
training.prediction=compute(nn,train.data[, -c(28:31,45)])
training.class1=apply(training.predictions$net.result,1,which.max)
a=as.factor(training.class1)
b=as.factor(train.nn$categori)
confusionMatrix(a,b)
```

Confusion Matrix and Statistics

		Reference			
Prediction	1	2	3	4	
1	224	36	13	3	
2	81	211	136	35	
3	10	7	24	19	
4	25	92	163	293	

Overall Statistics

Accuracy : 0.5614  
95% CI : (0.5213, 0.5747)

No Information Rate : 0.2551  
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.3847

McNemar's Test P-Value : < 2.2e-16

Statistics by Class:

	Class: 1	Class: 2	Class: 3	Class: 4
Sensitivity	0.6588	0.6204	0.07132	0.8371
Specificity	0.9436	0.7544	0.96541	0.7260

میزان درستی در این مدل 0.5614 است

## برازش مدل و ماتریس درهم ریختگی بر روی داده‌های اعتبارسنجی

*valid*

```
validation.prediction=compute(nn,valid.data[, -c(28:31,45)])
validation.class1=apply(validation.prediction$net.result,1,which.max)
confusionMatrix(as.factor(validation.class1),as.factor(valid.nn$categori))
```

Confusion Matrix and Statistics

		Reference			
Prediction	1	2	3	4	
1	130	55	17	8	
2	73	112	74	40	
3	8	8	11	8	
4	21	67	116	166	

Overall Statistics

Accuracy : 0.469  
95% CI : (0.4264, 0.4919)

No Information Rate : 0.2656

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.2864

McNemar's Test P-Value : < 2.2e-16

Statistics by Class:

	Class: 1	Class: 2	Class: 3	Class: 4
Sensitivity	0.5603	0.4901	0.0604	0.7477
Specificity	0.8829	0.7217	0.96557	0.7056

میزان درستی در این مدل 0.469 است با مقایسه میزان درستی در داده‌های آموزشی و داده‌های اعتبار سنجی متوجه بیش برآش در

این مدل می‌شویم

## تحلیل تشخیصی

در روش تحلیل تشخیصی<sup>۱۷</sup> ابتدا کتابخانه‌های مورد نیاز را لود کرده و متغیر‌های گسسته را به فاکتور تبدیل می‌کنیم و مدل را بر روی بخش‌های آموزشی و اعتبارسنجی را که در قسمت‌های قبل مشخص کردیم، برازش می‌دهیم.

```
Discreme library(tidyverse)
library(caret)
library(MASS)
library(ggplot2)
data.df=Data
theme_set(theme_classic())
train.dec=train.data[,-c(6,7,13,1,3,5,6,7,8,9,10,12:42)]
valid.dec=valid.data[,-c(6,7,13,1,3,5,6,7,8,9,10,12:42)]
```

و در قسمت بعد متغیر‌ها را استاندارد می‌کنیم

```
preproc.param <- preProcess(train.dec,method = c("center", "scale"))
Transform the data using the estimated parameters
train.transformed <- preproc.param %>% predict(train.dec)
test.transformed <- preproc.param %>% predict(valid.dec)
```

## برازش مدل تحلیل تشخیصی و ماتریس برهم ریختگی بر روی داده‌های آموزشی

*Fit the model*

```
model <- lda(categuri~, data = train.transformed)
model

Call:
lda(categuri ~ ., data = train.transformed)

Prior probabilities of groups:
-1.32795941640475 -0.449059117655442  0.429841181093869   1.30874147984318
      0.2529155          0.2521866          0.2259475          0.2689504

Group means:
              Tedad        Age       S.Z H_Khorakivadokhani H_N
oshidani     H_Pushak    H_Maskan
-1.32795941640475 -0.33800142  0.07266091 -0.355189870      -0.5001035 -0.0
02908433 -0.14387924 -0.4371994
-0.449059117655442 -0.06797275 -0.03598980 -0.060317284      -0.1964942 -0.0
15538443 -0.07920914 -0.1666079
 0.429841181093869  0.11684161 -0.09114346 -0.001812877      0.1317377 -0.0
09860318 -0.01019495  0.1132483
 1.30874147984318  0.28342592  0.04198809  0.392093921      0.5438596  0.0
25588688  0.21813793  0.4722156
                                H_mobleman   H_behdasht H_Hamloghl H_Ertebatat H_Tafrihat H
_Ghazayeamade H_kalavakhadamat
-1.32795941640475 -0.16865011 -0.139294527 -0.3193541  -0.5377063 -0.2158569
-0.043255580      -0.25045850
-0.449059117655442 -0.09259120 -0.120306864 -0.1339759  -0.2475746 -0.1196000
-0.106060769      -0.12665397
 0.429841181093869  0.07473153  0.006808032  0.0203104  0.1670780 -0.0116531
-0.002516716      0.03734991
 1.30874147984318  0.18263244  0.238078282  0.4088762  0.5974274  0.3249225
0.142240906      0.32290760

Coefficients of linear discriminants:
              LD1        LD2        LD3
Tedad      0.29458841 -0.30200632  0.092944986
Age        -0.03433053  0.60496246 -0.135043093
S.Z        0.28632546  0.09117452  0.788809545
H_Khorakivadokhani 0.37173068 -0.23082758  0.016962932
H_Noshidani -0.01320036  0.10717896  0.054121897
H_Pushak   -0.03399122  0.06353744  0.044288753
H_Maskan   0.33895828 -0.36517226 -0.126453527
H_mobleman 0.01998005 -0.11153158 -0.251504065
H_behdasht 0.06960987  0.31900775 -0.145046145
```

```

H_Hamloghl      0.28424188  0.24011006  0.340679081
H_Ertebatat    0.43063928 -0.03680016 -0.530347063
H_Tafrihat      0.17111653  0.45516910  0.128325298
H_Ghazayeamade -0.08478638  0.27293008 -0.429057661
H_kalavakhadamat -0.03461935  0.28381388 -0.005808894

```

Proportion of trace:

LD1	LD2	LD3
0.9601	0.0275	0.0124

```

predictions <- model %>% predict(train.transformed)
as.factor(predict(model)$class)

> predict.dec=factor(predict(model)$class,levels = c(-1.32403275189865, -0.42161
2018860922,  0.480808714176809,   1.38322944721454 ),labels = c(1,2,3,4))

> cat=as.factor(train$categuri)

> confusionMatrix(predict.dec,cat)

```

Confusion Matrix and Statistics

#### Reference

Prediction	1	2	3	4
1	226	116	64	20
2	75	131	89	63
3	39	76	123	81
4	10	40	73	148

Overall Statistics

Accuracy : 0.4864

95% CI : (0.4297, 0.4897)

No Information Rate : 0.2566

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.2731

Mcnemar's Test P-Value : 0.0003753

Statistics by Class:

	Class: 1	Class: 2	Class: 3	Class: 4
Sensitivity	0.6457	0.34491	0.35244	0.4891
Specificity	0.8043	0.77445	0.86841	0.8830

میزان درستی در این مدل 0.4864 است

## برازش مدل تحلیل تشخیصی و ماتریس برهم ریختگی بر روی داده‌های اعتبارسنجی

```
model <- lda(categuri~, data = test.transformed)
predictions <- model %>% predict(test.transformed)
as.factor(predict(model)$class)

predict.dec=factor(predict(model)$class,levels = c(-1.32403275189865, -0.4216120
18860922, 0.480808714176809, 1.38322944721454 ),labels = c(1,2,3,4))
cat=as.factor(valid$categuri)
confusionMatrix(predict.dec,cat)
```

Confusion Matrix and Statistics

Prediction		Reference			
		1	2	3	4
1	151	67	37	18	
2	53	111	64	46	
3	6	20	47	36	
4	12	39	57	151	

Overall Statistics

Accuracy : 0.5127  
 95% CI : (0.4698, 0.5356)  
 No Information Rate : 0.2743  
 P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.3422

Mcnemar's Test P-Value : 9.098e-10  
 Statistics by Class:

	Class: 1	Class: 2	Class: 3	Class: 4
Sensitivity	0.6802	0.4684	0.22927	0.6016
Specificity	0.8240	0.7596	0.91268	0.8373

میزان درستی در این مدل 0.5127 است

## روش k امین نزدیک ترین همسایه

از آنجایی در این روش مبنای انجام کار فاصله است، مقیاس متغیرهای مختلف از جمله سن که بر حسب سال است، سطح زیربنای ساختمان که بر اساس مترمربع است و یا هزینه خوراک و پوشак که به ریال است باید یکسان شود. از این رو تمام متغیرهای پیوسته را استاندارد می‌کنیم تا واحد آن از بین بود و همچنین متغیرهای گسسته را فاکتور می‌کنیم که دستورات آن به صورت زیر است:

```
#knn  
new.data=final  
new.data[,1:27] <- lapply(new.data[,1:27] , factor)  
  
attach(new.data)  
set.seed(1234)  
train.row=sample(rownames(new.data),dim(new.data)[1]*0.5)  
valid.row=sample(setdiff(rownames(new.data),train.row),dim(new.data)[1]*0.35)  
test.row=setdiff(rownames(new.data),union(train.row,valid.row))  
train.data=new.data[train.row,]  
valid.data=new.data[valid.row,]  
test.data=new.data[test.row,]  
  
train.norm.df <- train.data  
valid.norm.df <- valid.data  
mower.norm.df <- test.data  
# use preProcess() from the caret package to normalize Income and Lot_Size.  
library(caret)  
library(FNN)  
norm.values <- preProcess(train.data[, 28:39], method=c("center", "scale"))  
train.norm.df[, 28:39] <- predict(norm.values, train.data[, 28:39])  
valid.norm.df[, 28:39] <- predict(norm.values, valid.data[, 28:39])
```

```
mower.norm.df[, 28:39] <- predict(norm.values, test.data[, 28:39])
```

در این قسمت بهترین  $k$  را انتخاب میکنیم

```
accuracy.df <- data.frame(k = seq(1, 30, 1), accuracy = rep(0, 30))
cc=as.factor(train.norm.df$categori)
c=as.factor(valid.norm.df$categori)
for(i in 1:30) {
  knn.pred <- knn(train.norm.df[, 28:39], valid.norm.df[, 28:39], cl =cc, k = i,prob = TRUE)
  accuracy.df[i, 2] <- confusionMatrix(knn.pred,c)$overall[1]
}
accuracy.df
   k accuracy
1  1  0.39125
2  2  0.36500
3  3  0.36875
4  4  0.37375
5  5  0.37125
6  6  0.38250
7  7  0.40375
8  8  0.39375
9  9  0.40375
10 10  0.41375
11 11  0.42375
12 12  0.42125
13 13  0.42750
14 14  0.43000
15 15  0.42500
16 16  0.42375
17 17  0.42750
18 18  0.41250
19 19  0.41500
20 20  0.42000
21 21  0.42000
22 22  0.40625
23 23  0.41375
24 24  0.40750
25 25  0.40375
26 26  0.41875
27 27  0.40000
28 28  0.40750
29 29  0.41500
30 30  0.41875
```

که مشخص می شود بهترین درستی متعلق به  $k=14$  است

## مدل بر روی داده‌های اعتبارسنجی

```
class = as.factor(train.norm.df$categori)
cl=as.factor(valid.norm.df$categori)
nn <- knn(train = train.norm.df[, 28:39], test = valid.norm.df[,28:39], class, k =14 )
confusionMatrix(nn,cl)
Confusion Matrix and Statistics
```

Prediction	Reference			
	1	2	3	4
1	139	63	39	12
2	37	76	51	46
3	24	44	58	69
4	13	23	35	71

Overall statistics

Accuracy : 0.43  
95% CI : (0.3954, 0.4651)  
No Information Rate : 0.2662  
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.2385

McNemar's Test P-Value : 4.543e-05

Statistics by Class:

	Class: 1	Class: 2	Class: 3	Class: 4
Sensitivity	0.6526	0.3689	0.3169	0.35859
Specificity	0.8058	0.7744	0.7780	0.88206

مقدار درستی برابر 0.43 شد و مدل در تشخیص رده 1 درآمدی عملکرد بهتری دارد و همانطور که انتظار داشتیم اعداد مشخص سازی

برای مدل بیشتری از حساسیت است زیرا عدم تشخیص رده ها بیشتر از تشخیص درست رده ها است.

## مدل بر روی داده های آزمون

از آنجایی که در تعیین  $k$  از داده های اعتبار سنجی استفاده کردیم برای ارزیابی مدل باید از داده های تست استفاده کنیم

```
nn.t <- knn(train = train.norm.df[, 28:39], test = mower.norm.df[,28:39], class, k = 7)
```

```
t=as.factor(mower.norm.df$categori)
```

```
confusionMatrix(nn.t,t)
```

Confusion Matrix and Statistics

		Reference			
		1	2	3	4
Prediction	1	61	30	11	13
	2	14	31	30	17
		3	11	20	27
		4	3	13	12

Overall Statistics

Accuracy : 0.4331

95% CI : (0.3801, 0.4873)

No Information Rate : 0.2733

P-value [Acc > NIR] : 1.436e-10

Kappa : 0.2415

McNemar's Test P-value : 0.009079

Statistics by Class:

	Class: 1	Class: 2	Class: 3	Class: 4
Sensitivity	0.6854	0.32979	0.33750	0.37037
Specificity	0.7882	0.75600	0.80303	0.89354

همانطور که می بینیم مدل بر روی داده های آزمون نتیجه مشابه داده های اعتبار سنجی را داده است حتی مقدار درستی 0.0031

افزایش داشته است

## تفسیر و بررسی نتایج و برازش بهترین مدل

	درستی داده‌های اعتبارسنجی	درستی داده‌های آموزشی
درخت رده بندی	<b>0.459</b>	<b>0.4</b>
K-nn	<b>0.43</b>	<b>0.433</b>
رگرسیون ترتیبی	<b>0.474</b>	<b>0.475</b>
شبکه عصبی	<b>0.5614</b>	<b>0.469</b>
SVM	<b>0.55</b>	<b>0.57</b>
تحلیل تشخیصی	<b>0.486</b>	<b>0.512</b>

بهترین مدل، مدل SVM با درستی 0.55 برای داده‌های آموزشی و 0.57 برای داده‌های اعتبار سنجی است.

## منابع

انگلیسی

1. DATA MINING FOR BUSINESS ANALYTICS ,Galit Shmueli Peter C. Bruce Inbal Yahav Nitin R. Patel Kenneth C. Lichtendahl, Jr
2. ggplot2 Elegant Graphics for Data Analysis , Hadley Wickham
3. Guide to Create Beautiful Graphics in R, Alboukadel Kassambara
4. <https://fa.wikipedia.org/wiki/>

فارسی

کتاب تحقیقات بازار در یه هفته از پالی برد

## پیوست

### دستورات بخش تصویری سازی

```
data

library(readxl)

Data <- read_excel("C:/Users/Administrator/Desktop/Data mining final project/Data.xlsx")

data.df=orginal

data.df=data.frame(Data)

attach(data.df)

Agegroup=cut(Age,breaks=c(18,25,35,45,55,91), labels=c("18-25","26-35", "36-45","46-55","56-91"))

data.df=cbind(Agegroup,data.df)

View(data.df)

dpermonth=rowSums(data.df[,65:68])

Hazine=rowSums(data.df[,54:64])

data.df=cbind(data.df,dpermonth,Hazine)

q=quantile(dpermonth,.7)

response=ifelse(dpermonth>q,1,0)

data.df=cbind(data.df,response)

View(data.df)

data.df=data.df[,-c(2,3)]

names(data.df)
```

```
attach(data.df)

library(ggplot2)

View(data.df)

a<-ggplot(data=df, aes(x =oto))

a + geom_bar()

pcs <- prcomp(data=df,scale. = T)

summary(pcs)

pcs$rot

scores <- pcs$x

head(scores,10)

gender=as.factor(Gender)

a<-ggplot(data=df, aes(x =log(dpermonth)))

a + geom_bar(stat="bin")

a + geom_histogram(aes(color = gender, fill = gender),

alpha = 0.4, position = "identity") + 

scale_fill_manual(values=c("00AFBB", "E7B800")) + 

scale_color_manual(values=c("00AFBB", "E7B800"))
```

```
age

hist(Age)

plot(density(Age), col="brown1")

summary(Age)

Agegroup=cut(Age,breaks=c(18,25,35,45,55,91), labels=c("18-25","26-35", "36-45","46-55","56-91"))

table(Agegroup)

summary(Agegroup)

plot(Agegroup)

library(ggplot2)

data.plot=data.frame(table(Agegroup))

library(RColorBrewer)

p <- ggplot(data=data.df, aes(x=Agegroup), fill=brown1)

p + geom_bar(col="black",width=.5, fill="cyan3")



data.plot=data.frame(table(Agegroup))

gender

x=table(Gender)

library(expss)
```

```

val_lab(gender) = num_lab("

  2 woman

  1 man

")

gender=as.character(Gender)

library(ggplot2)

use_labels(mtcars, {

  p <- ggplot(data=data.df, aes(x=gender), fill=brown1)

  p + geom_bar(col="black",width=.4, fill="coral1")

})

man=Gender==1

p <- ggplot(data=data.df, aes(x=Agegroup, fill=man))

p + geom_bar(col="red")

tedad

table(Tedad.a)

number_of_family=as.character(Tedad.a)

as.factor(Tedad.a)

p <- ggplot(data=data.df, aes(x=number_of_family), fill=brown1)

p + geom_bar(col="black",width=.6, fill="cyan3")

oto

```

```
p <- ggplot(data=data.df, aes(x=oto), fill=brown1)

p + geom_bar(col="black",width=.5, fill="coral1")

table(oto)

savad

table(Savad)

val_lab(Savad) = num_lab("

  2 No

  1 Yes

")

Savad=as.character(Savad)

class(Savad)

table(Savad)

use_labels(mtcars, {

  p <- ggplot(data=data.df, aes(x=Savad), fill=brown1)

  p + geom_bar(col="black",width=.5, fill="cyan3")

})

inedu

table(InEdu)
```

```

education=as.character(InEdu)

p <- ggplot(data=data.df, aes(x=InEdu), fill=brown1)

p + geom_bar(col="black",width=.5, fill="cyan3")

madrak

table(n.t.m)

table(Madrak)

library(expss)

is.na.data.frame(data.df)

Madrak.factor=as.factor(Madrak)

na.omit(Madrak.factor)

table(Faaliat)

table(Madrak)

p <- ggplot(data=data.df, aes(x=Madrak), fill=brown1)

p + geom_bar(col="black",width=.6, fill="cyan3") + scale_x_discrete(limits =
c("ebtedai","rahnmai","motevasete",
  "diplom","foghdiplom","lisans","arshad",
  "doktora","bisavad"))

```

```

faaliat

table(Faaliat)

Faaliat=as.factor(Faaliat)

p <- ggplot(data=data.df, aes(x=Faaliat), fill=brown1)

p + geom_bar(col="black",width=.5, fill="cyan3")+scale_x_discrete(limits =
c("shaghel","bikar(joya)","badpermonth bikar","mohasel","khanedar","sayer" ))


library(ggplot2)

ss <- ggplot(data=data.df, aes(x=Faaliat, fill=brown1))

ss + geom_bar(col="black",width=.6, fill="cyan3") + scale_x_discrete(limits =
c("shaghel","bikar(joya)","badpermonth bikar","sayer" ))


table(Faaliat)

View(data.df)

library(forcats)

library(dplyr)

replace(Faaliat, list=c(2,5,6), values=c(4,4,4))

Faaliat

table(Faaliat)

library(plotrix)

```

```

slice=c(351,7,113,4,5)

x1="Ø Ø§ØºÙº"
x2="ØºÙºÚºØ§Ø± (Ø¬ÙºÙºØ§Ùº ÚºØ§Ø±"
x3="ØºØ§Ø±Ø§Ùº ØºØ±Ø§ÙºØº ØºØºÙº ÚºØ§Ø±"
x4="ØºØ§ÙºÙº ØºØ§Ø±"
x5="Ø³Ø§ÙºØ±"

lbls=c(x1,x2,x3,x4,x5)

df=data.frame(slice,lbls)

library(ggplot2)

bp<- ggplot(df, aes(x="", y=slice, fill=lbls))+

  geom_bar(width = 1, stat = "identity")



pie <- bp + coord_polar("y", start=0)

pie

t.shaghel

table(T.shaghel)

p <- ggplot(data=data.df, aes(x=T.shaghel), fill=brown1)

p + geom_bar(col="black",width=.5, fill="cyan3") + scale_x_discrete(limits = c("1","2","3","4" ))

T.O

table(T.O)

p <- ggplot(data=data.df, aes(x=T.O), fill=brown1)

```

```
p      + geom_bar(col="black",width=.5,      fill="cyan3")+scale_x_discrete(limits      =
c("2","3","4","5","6","8"))
```

N.S

```
table(N.S)
```

```
p <- ggplot(data=data.df, aes(x=N.S), fill=brown1)
```

```
p      + geom_bar(col="black",width=.5,      fill="coral1")+scale_x_discrete(limits      =
c("felezi","botoni","sayer"))
```

masaleh

```
table(Masleh)
```

```
p <- ggplot(data=data.df, aes(x=Masleh), fill=brown1)
```

```
p      + geom_bar(col="black",width=.5,      fill="cyan3")+scale_x_discrete(limits      =
c("Ahan","Ajor","siman","kheshti"))
```

ntm

```
table(n.t.m)
```

```
p + geom_bar(col="black",width=.6, fill="coral1") + scale_x_discrete(limits = c("malek  
kol","malek","ejari","rahn","khedmat","free"))
```

```
boxplot( Age~ cat , ylab = "age", xlab = "dpermonth", )
```

```
group=as.factor(cat)
```

```
p1 <- ggplot(data.df, aes(x=group,y=Age,color=group))
```

```
p1 + geom_boxplot(outlier.size=0)
```

```
savad=as.factor(savadd)
```

```
p2 <- ggplot(data.df, aes(x=savad,y=dpermonth))
```

```
p2 + geom_boxplot(outlier.size=0)
```

```
p3 <- ggplot(data.df, aes(x=group,y=H_Khorakivadokhani,color=group))
```

```
p3 + geom_boxplot(outlier.size=0)
```

```
p3 <- ggplot(data.df, aes(x=group,y=H_Noshidani,color=group))
```

```
p3 + geom_boxplot(outlier.size=0)+ylim(0,1000000)
```

```
p3 <- ggplot(data.df, aes(x=group,y=H_behdasht,color=group))
```

```
p3 + geom_boxplot(outlier.size=0)
```

```
p4 <- ggplot(data.df, aes(x=group,y=H_Maskan,color=group))
```

```
p4 + geom_boxplot(outlier.size=0)

p5 <- ggplot(data.df, aes(x=group,y=H_Hamloghl,color=group))
p5 + geom_boxplot(outlier.size=0)

p6 <- ggplot(data.df, aes(x=group,y=S.Z,color=group))
p6 + geom_boxplot(outlier.size=0)

p8 <- ggplot(data.df, aes(x=group,y=H_kalavakhadamat,color=group))
p8 + geom_boxplot(outlier.size=0)

p9 <- ggplot(data.df, aes(x=group,y=H_Pushak,color=group))
p9 + geom_boxplot(outlier.size=0)

ggarrange(p1,p2,p3,p4,p5,p6,p7,p8,p9)

+geom_jitter(position= position_jitter(h=.1))

par(mfrow=c(2,2))

library(ggplot2)

gender=as.factor(genderr)
```

```
p <- ggplot(data.df, aes(x=gender, y=log(dpermonth)), color=gender) +  
  geom_boxplot(color="cyan3", fill="cyan", alpha=0.2)
```

p

```
r<- ggplot(data.df, aes(x=Agegroup, y=log(dpermonth)), fill=Agegroup) +  
  geom_boxplot(color="cyan3", fill="cyan",alpha=0.2)
```

r

```
manzel=as.factor(T.M.S.)
```

```
t<- ggplot(data.df, aes(x=manzel, y=dpermonth), fill=manzel) +  
  geom_boxplot(color="coral", fill="coral1",alpha=0.2)
```

t

```
N.S.plot=as.factor(N.SS)
```

```
r<- ggplot(data.df, aes(x=N.S.plot, y=dpermonth), fill=N.S.plot) +  
  geom_boxplot(color="cyan3", fill="cyan",alpha=0.2)
```

r

```
m=na.omit(Masleh)
```

```
masaleh=as.factor(Maslehh)
```

```
t<- ggplot(data.df, aes(x=masaleh, y=dpermonth), fill=masaleh) +
```

```
geom_boxplot(color="coral", fill="coral1",alpha=0.2)

t

faaliat.factor=as.factor(Faaliatt)

tabl

w<- ggplot(data=df, aes(x=faaliat.factor, y=dpermonth), fill=faaliat.factor) +
  geom_boxplot(color="cyan3", fill="cyan",alpha=0.2)

w

na.omit(Madrak.factor)

Madrak.factor=as.factor(madrakk)

z<- ggplot(data=df, aes(x=Madrak.factor, y=log(dpermonth)), fill=Madrak.factor) +
  geom_boxplot(color="cyan3", fill="cyan",alpha=0.2)

z

table(Madrak)

ggarrange(p,r)

library(ggpubr)

library(expss)

val_lab(manzel) = num_lab("

  6 free
```

5 khedmat

4 rahn

3 ejari

2 malek

1 malekkol

")

```
manzel=as.character(manzel)
```

```
hazine.log=log(Hazine);dpermonth.log=log(dpermonth)
```

S.Z

```
p <- ggplot(data=data.df, aes(x=dpermonth.log, y=hazine.log))
```

```
p + geom_point(size=4)
```

```
p <- ggplot(data=data.df, aes(x=S.Z))
```

```
p + geom_histogram()
```

hazine

```
H_Taghzie=rowSums(data.df[,c(54,55,63)])
```

```
data.df=cbind(data.df,H_Taghzie)
```

```
View(data.df)
```

```
summary(H_Taghzie)
```

```
na.omit(H_Taghzie)
```

```
na.omit(H_Pushak)
```

```
na.omit(H_Maskan)
```

```
na.omit(H_behdasht)
```

```
na.omit(H_Hamloghl)
```

```
na.omit(H_Ertebatat)
```

```
na.omit(H_Tafrihat)
```

```
na.omit(H_kalavakhadamat)
```

```
a <- ggplot(data=data.df, aes(x=H_Khorakivadokhani))
```

```
a + geom_histogram()
```

```
a <- ggplot(data=data.df, aes(x=H_Noshidani))
```

```
a + geom_histogram()
```

```
a <- ggplot(data=data.df, aes(x=H_Ghazayeamaade))
```

```
a + geom_histogram()
```

```
a <- ggplot(data=data.df, aes(x=H_mobleman))
```

```
a + geom_histogram()
```

```
b <- ggplot(data=data.df, aes(x=H_Pushak))
```

```
b + geom_histogram(bins=55)
```

```
c <- ggplot(data=data.df, aes(x=H_Maskan))
```

```
c + geom_histogram(bins=55)
```

```
d <- ggplot(data=data.df, aes(x=H_behdasht))
```

```
d + geom_histogram(bins=55)
```

```
e <- ggplot(data=data.df, aes(x=H_Hamloghl))
```

```
e + geom_histogram(bins=55)
```

```
f <- ggplot(data=data.df, aes(x=H_Ertebatat))
```

```
f + geom_histogram(bins=55)
```

```
g <- ggplot(data=data.df, aes(x=H_Tafrihat))
```

```
g + geom_histogram(bins=55)
```

```
h <- ggplot(data=data.df, aes(x=H_kalavakhadamat))
```

```

h + geom_histogram(bins=55)

ggarrange(c,e,f,h)

?stat_bin


g <- ggplot(data=data.df, aes(x=dpermonth))

g + geom_histogram()

na.omit(H_Pushak)

meanhazineha=colMeans(data.df[,c(56,57,58,59,60,61,62,64)])


H.means=colMeans(data.df[,c(56,57,58,59,60,61,62,64,69)]) 

barplot(H.means, col="coral1", main="Hazine",density =100,width = .9)


D.means=colMeans(data.df[,c(65,66,67,68)]) 

barplot(D.means, col="cyan3", main="dpermonth",density =100,width = .9)


hararati

library(gplots)

heatmap.2(cor(data.df[,c(4,11,43:53,59)]),Rowv = F,Colv = F,dendrogram = "none",cellnote = round(cor(data.df[,c(4,11,43:53,59)]),2),notecol = "black",key = F,trace = "none",margins = c(10,10),
main = "نماذج حرارتی")

cate=factor(cat,levels = c("A","B","C","D"),labels = c("1","2","3","4"))

```

matrici

```
plot(data.df[,c(4,11,43:53,59)])
```

```
library(GGally)
```

```
ggpairs(data.df[,c(43:53,59)])
```

```
par(xpd=TRUE) # allow legend to be displayed outside of plot area
```

```
plot(housing.df$NOX ~ housing.df$LSTAT, ylab = "NOX", xlab = "LSTAT",
```

```
  col = ifelse(housing.df$CAT..MEDV == 1, "black", "gray"))
```

```
library(ggplot2)
```

3D

```
ggplot(data.df, aes(y = H_Maskan, x = S.Z, colour = response)) + geom_point(alpha = 0.6)
```

```
ggplot(data.df, aes(y = H_Khorakivadokhani, x = H_kalavakhadamat, colour = response)) +  
  geom_point(alpha = 0.6)
```

```
ggplot(data.df, aes(y = H_behdasht, x = H_Tafrihat, colour = response)) + geom_point(alpha = 0.6)
```

```
ggplot(data.df, aes(y = H_behdasht, x = H_Ertebatat, colour = response)) + geom_point(alpha = 0.6)
```

```
ggplot(data.df, aes(y = H_Maskan, x = H_kalavakhadamat, colour = response)) + geom_point(alpha = 0.6)
```

```
ggplot(data=df, aes(y = H_Maskan, x =H_Khorakivadokhani, colour= response)) + geom_point(alpha = 0.6)
```

scatter plot

```
p <- ggplot(data=data.df, aes(x=Age, y=dpermonth))

p + geom_point(shape=1) + geom_smooth()
```

```
p <- ggplot(data=data.df, aes(x=H_Maskan, y=dpermonth))
```

```
p + geom_point(shape=1) + geom_smooth()
```

```
p <- ggplot(data=data.df, aes(x=H_behdasht, y=dpermonth))
```

```
p + geom_point(shape=1) + geom_smooth()
```

```
p <- ggplot(data=data.df, aes(x=H_Ertebatat, y=dpermonth))
```

```
p + geom_point(shape=1) + geom_smooth()
```

```
p <- ggplot(data=data.df, aes(x=H_Hamloghl, y=dpermonth))
```

```
p + geom_point(shape=1) + geom_smooth()
```

```
boxplot(H_Hamloghl)
```

```
boxplot(y)
```

```
p <- ggplot(data=data.df, aes(x=H_Tafrihat, y=dpermonth))
```

```
p + geom_point(shape=1) + geom_smooth()
```

```

p <- ggplot(data=data.df, aes(x=H_mobleman, y=dpermonth))

p + geom_point(shape=1) + geom_smooth()

p <- ggplot(data=data.df, aes(x=H_Khorakivadokhani, y=dpermonth))

p + geom_point(shape=1) + geom_smooth()

p <- ggplot(data=data.df, aes(x=H_behdasht, y=dpermonth))

p + geom_point(shape=1) + geom_smooth()

```

hararati

```

library(gplots)

colMeans(x)

x=na.omit(data.df[,c(1,2,3,6:11,13,14,61)])

heatmap.2(cor(x),Rowv = F,Colv = F,dendrogram = "none",cellnote = round(cor(x),2),notecol =
"black",key = F,trace = "none",margins = c(10,10), main = "نمودار حرارتی")

data.df>Data

attach(data.df)

boxplot(dpermonth~oto)

```

```
library(ggplot2)

p <- ggplot(data=data.df, aes(x=otoo, y=dpermonth))

p + geom_boxplot(shape=1) + geom_smooth()

car=as.factor(oto)
```

## دستورات بخش درخت رگرسیون

```
data.df=Data

tree

data.df[,c(1,3,5,6,7,8,9,10,12:42)] <- lapply(data.df[,c(1,3,5,6,7,8,9,10,12:42)], factor)

View(data.df)

library(rpart)

library(rpart.plot)

library(caret)

library(lattice)

library(ggplot2)

library(e1071)

data.df=Data

train.index<-sample(c(1:dim(data.df)[1]),dim(data.df)[1]*0.6)

train<-data.df[train.index,]

valid<-data.df[-train.index,]
```

```
train.tree=train[,-c(54:59,6,7,13)]  
  
valid.tree=valid[,-c(54:59,6,7,13)]  
  
attach(data.df)  
  
class.tree <- rpart( categori~ ., data = train.tree, method = "class",model=TRUE)  
  
prp(class.tree, type = 1, extra = 1, under = TRUE, split.font = 2, varlen = -10,  
cex=.9,box.palette=c("mediumvioletred"))
```

```
class.tree.t <- predict(class.tree,train.tree,type = "class")  
  
confusionMatrix(as.factor(class.tree.t),as.factor(train$categori))
```

```
class.tree.v<-rpart(categor~., data = valid.tree ,method = "class", model=TRUE)  
  
prp(class.tree.v,type =0,extra = 1,under = T,split.font = 1,  
varlen = -10,cex=.7,box.palette=c("mediumvioletred"))
```

```
class.tree.p <- predict(class.tree.v,valid.tree,type = "class")  
  
confusionMatrix(as.factor(class.tree.p),as.factor(valid$categori))
```

deeper tree

train

```

deeper.ct <- rpart(categuri ~ ., data = train.tree, method = "class", cp = 0, minsplit = 1)

length(deeper.ct$frame$var[deeper.ct$frame$var == "<leaf>"])

prp(deeper.ct, type = 1, extra = 1, under = TRUE, split.font = 1, varlen = -10,
     box.col=ifelse(deeper.ct$frame$var == "<leaf>", 'gray', 'white'))

class.tree.p <- predict(deeper.ct,train.tree,type = "class")

confusionMatrix(as.factor(class.tree.p),as.factor(train.tree$categuri))

deeper.ct <- rpart(categuri ~ ., data = train.tree, method = "class", cp = 0, minsplit = 1)

deeper tree

validation

deeper.ct.v <- rpart(categuri ~ ., data = valid.tree, method = "class", cp = 0, minsplit = 1)

class.tree.v <- predict(deeper.ct.v,valid.tree,type = "class")

confusionMatrix(as.factor(class.tree.v),as.factor(valid.tree$categuri))

deeper.ct <- rpart(categuri ~ ., data = train.tree, method = "class", cp = 0, minsplit = 1)

deeper.tree<- rpart(categuri ~ ., data = train.tree, method = "class",cp= .00001,minsplit =5)

options(digits=8)

printcp(deeper.tree)

which.min(deeper.tree$cptable[, "xerror"])

```

```
library(rpart)

cv.tree<-rpart(categuri~,data = train.tree ,method = "class",
               cp = 0.003917728, minsplit =5,xval =5, model=T)

prp(cv.tree,type =0,extra = 1,under = T,split.font = 2,
     varlen = -10,cex=.7,box.palette=c("mediumvioletred"))

cv.tree.p <- predict(cv.tree,train.tree,type = "class")

confusionMatrix(as.factor(cv.tree.p),as.factor(train$categuri))

cv.tree.v<-rpart(categuri ~., data = valid.tree,method = "class",cp=0.003917728 , minsplit =5,xval =5,
model=TRUE)

prp(cv.tree.v,type =0,extra = 2,under = T,split.font = 2,
     varlen = -8,cex=.5,box.palette=c("mediumvioletred"))

cv.tree.p.v <- predict(cv.tree.v,valid,type = "class")

confusionMatrix(as.factor(cv.tree.p.v),as.factor(valid$categuri))
```

## SVM دستورات بخش

```
data.df=Data

data.df[,c(1,3,5,6,7,8,9,10,12:42)] <- lapply(data.df[,c(1,3,5,6,7,8,9,10,12:42)], factor)

train.index<-sample(c(1:dim(data.df)[1]),dim(data.df)[1]*0.6)

train<-data.df[train.index,]

valid<-data.df[-train.index,]

train.svm=train[,-c(54:59,6,7,13,1,3,5,6,7,8,9,10,12:42)]

valid.svm=valid[,-c(54:59,6,7,13,1,3,5,6,7,8,9,10,12:42)]

library(e1071)

library(caret)

attach(data.df)

View(train.svm)

dim(train.svm)

model=svm(categuri~,data = train.svm,type="C-classification")

summary(model)

View(train.svm)

x=train.svm[,-15]

pred <- predict(model,x )

cate=as.factor(train.svm$categuri)

confusionMatrix(pred,cate)

categuri=as.factor(categuri)

valid
```

```

model.v=svm(categuri~,data = valid.svm,type="C-classification")

summary(model)

pred <- predict(model.v, x1)

confusionMatrix(pred,categ)

categ=as.factor(valid.svm$categuri)

x1=valid.svm[,-15]

pred <- predict(model.v,x )

cate=as.factor(train.svm$categuri)

confusionMatrix(pred,cate)

categuri=as.factor(categuri)

```

## دستورات بخش KNN

```

knn

categuri=as.factor(categuri)

train.index<-sample(c(1:dim(knn)[1]),dim(knn)[1]*0.6)

train.knn<-knn[train.index,]

valid.knn<-knn[-train.index,]

cc=cl1[!is.na(cl1)]

library(caret)

norm.values <- preProcess(train.knn[,37:56], method=c("center", "scale"))

```

```
norm.values.v <- preProcess(valid.knn[,37:56], method=c("center", "scale"))
```

```
train.knn[, 37:56] <- predict(norm.values, train.knn[, 37:56])
```

```
valid.knn[, 37:56] <- predict(norm.values.v, valid.knn[, 37:56])
```

training

```
knn <- knn(train = train.knn[, 37:56], test = train.knn[, 37:56],cl=cc , k = 7)
```

```
k=as.factor(knn)
```

```
b=as.factor(train.knn$categori)
```

```
confusionMatrix(k,b)
```

validation

```
knn <- knn(train = train.knn[, 37:56], test = valid.knn[, 37:56],cl=cc , k = 7)
```

```
k=as.factor(knn)
```

```
b=as.factor(valid.knn$categori)
```

```
confusionMatrix(k,b)
```

## دستورات بخش شبکه عصبی

NN

attach(data)

summary(Age)

Agegroup=cut(Age,breaks=c(18,25,35,45,55,100), labels=c("18-25","26-35", "36-45","46-55","56-100"))

dato=cbind(Agegroup,data)

Gender=as.factor(Gender)

Ostan=as.factor(Ostan)

Agegroup=as.factor(Agegroup)

Faaliat=as.factor(Faaliat)

Madراك=as.factor(Madراك)

oto=as.factor(oto)

Pc=as.factor(Pc)

mobile=as.factor(mobile)

daramad.class=as.factor(categuri)

x1=class.ind(Gender)

x2=class.ind(Ostan)

x3=class.ind(mobile)

x4=class.ind(Faaliat)

x5=class.ind(Madراك)

x6=class.ind(oto)

```

x7=class.ind(Pc)

x8=class.ind(daramad.class)

dums=data.frame(x1,x2,x3,x4,x5,x6,x7,x8)

colnames(dums)=c(paste("gender",c(1,2),sep=""),
                 paste("Ostan",c(1,2,3,4,5),sep=""),
                 paste("mobile",c(0,1),sep=""),
                 paste("faaliat",c(1,2,3,4,5,6),sep=""),
                 paste("madrak",c(1,4,5,6,7,8,9,10),sep=""),
                 paste("car",c(0,1),sep=""),
                 paste("pc",c(0,1),sep=""),
                 paste("daramad.class",c(1,2,3,4),sep=""))

model=lm(log(daramad)~Gender+Ostan+Faaliat+Madrak+oto+Tedad+T.O,data = dums)

summary(model)

norm.values <- preProcess(dato[,c(2,9,40:50)], method=c("center", "scale"))

dato[,c(2,9,40:50)] <- predict(norm.values, dato[,c(2,9,40:50)])

data.nn=cbind(dums,dato[,c(2,9,40:50,57)])

train.index<-sample(c(1:dim(data.nn)[1]),dim(data.nn)[1]*0.6)

train.nn<-data.nn[train.index,]

valid.nn<-data.nn[-train.index,]

library(MASS)

library(lattice)

```

```

library(ggplot2)

library(e1071)

library(caret)

library(rpart)

library(rpart.plot)

library(nnet)

library(neuralnet)

nn=neuralnet(daramad.class1+daramad.class2+daramad.class3+daramad.class4~gender1
+gender2+
Ostan1+
Ostan2+
Ostan3+
Ostan4+
Ostan5+mobile0+
mobile1+
faaliat1+
faaliat2+
faaliat3+
faaliat4
+faaliat5
+faaliat6
+madrak1
+madrak4
+madrak5
+madrak6+
madrak7+
madrak8+
madrak9+
madrak10+
car0
+car1+
pc0+
pc1+
Tedad+
S.Z+
H_Khorakivadokhani
+H_Noshidani
+H_Pushak+
H_Maskan+
H_mobleman+
H_behdasht+
H_Hamloghl
+H_Ertebatat+
H_Tafrihat+
H_Ghazayeamaade
+H_kalavakhadamat
,data = train.nn, linear.output = F, hidden =c(3,2))

plot(nn,rep="best")

train

training.prediction=compute(nn,train.nn[,-c(28:31,45)])

training.class1=apply(training.prediction$net.result,1,which.max)

a=as.factor(training.class1)

b=as.factor(train.nn$categori)

confusionMatrix(a,b)

```

valid

```
validation.prediction=compute(nn,valid.nn[,-c(28:31,45)])  
  
validation.class1=apply(validation.prediction$net.result,1,which.max)  
  
confusionMatrix(as.factor(validation.class1),as.factor(valid.nn$categori))
```

## دستورات بخش رگرسیون ترتیبی

reg

```
attach(Data)  
  
dato$Ostan=factor(Ostan,levels = c(1,2,3,4,5),labels =c("Tehran","Esfahan","Mashhad","Shiraz","Tabriz"))  
  
dato$Gender=factor(Gender,levels = c(1,2),labels =c("Male","Female"))  
  
dato$Savad=factor(Savad,levels = c(1,2),labels =c("Basavad","Bisavad"))  
  
dato$InEdu=factor(InEdu,levels = c(1,2),labels =c("yes","No"))  
  
dato$Madrak=factor(Madrak,levels = c(1,4,5,6,7,8,9),labels =c("zirdiplom","Diplom","Kardani","Karshenasi","Arshad","Doctora","sayer"))  
  
dato$Faaliat=factor(Faaliat,levels = c(1,2,3,4,5,6),labels =c("Shaghel","Bikar","Bikarbadaramad","mohasel","khanedar","sayer"))  
  
dato$N.S=factor(N.S,levels = c(1,2,3),labels =c("Felezi","Betoni","Sayer"))  
  
dato$Masleh=factor(Masleh,levels = c(1,2,3,4,5,6,7),labels =c("Ajor&Ahan&sang","Ajor&Chob&sang","blocksimani","Ajor","Chob","khesht","sayer"))  
  
dato$sookht=factor(sookht,levels = c(3,4),labels =c("gazmaye","gaztabii"))  
  
dato$oto=factor(oto,levels = c(0,1),labels = c("No","Yes"))  
  
dato$motor=factor(motor,levels = c(0,1),labels = c("No","Yes"))  
  
dato$do=factor(do,levels = c(0,1),labels = c("No","Yes"))
```

```
dato$radio=factor(radio,levels = c(0,1),labels = c("No","Yes"))

dato$zabt=factor(zabt,levels = c(0,1),labels = c("No","Yes"))

dato$TV=factor(TV,levels = c(0,1),labels = c("No","Yes"))

dato$DVD=factor(DVD,levels = c(0,1),labels = c("No","Yes"))

dato$Pc=factor(Pc,levels = c(0,1),labels = c("No","Yes"))

dato$mobile=factor(mobile,levels = c(0,1),labels = c("No","Yes"))

dato$yakhchal.f=factor(yakhchal.f,levels = c(0,1),labels = c("No","Yes"))

dato$gaz=factor(gaz,levels = c(0,1),labels = c("No","Yes"))

dato$jaro.b=factor(jaro.b,levels = c(0,1),labels = c("No","Yes"))

dato$m.lebas=factor(m.lebas,levels = c(0,1),labels = c("No","Yes"))

dato$charkh.kh=factor(charkh.kh,levels = c(0,1),labels = c("No","Yes"))

dato$panke=factor(panke,levels = c(0,1),labels = c("No","Yes"))

Data$cooler=factor(cooler,levels = c(0,1),labels = c("No","Yes"))

Data$m.zarf=factor(m.zarf,levels = c(0,1),labels = c("No","Yes"))

Data$microfer=factor(microfer,levels = c(0,1),labels = c("No","Yes"))

Data$bargh=factor(bargh,levels = c(0,1),labels = c("No","Yes"))

Data$tel=factor(tel,levels = c(0,1),labels = c("No","Yes"))

Data$internet=factor(internet,levels = c(0,1),labels = c("No","Yes"))

Data$hamam=factor(hamam,levels = c(0,1),labels = c("No","Yes"))

Data$hararat.m=factor(hararat.m,levels = c(0,1),labels = c("No","Yes"))

Data$package=factor(package,levels = c(0,1),labels = c("No","Yes"))

Data$fazelab=factor(fazelab,levels = c(0,1),labels = c("No","Yes"))
```

```

Data$T.M.S=factor(T.M.S,levels           =           c(1,3,4,5,6,7),labels           =
c("Malek","ejari","Rahn","Khedmat","Free","sayer"))

Data$categuri=as.factor(Data$categuri)

dato$internet=as.factor(dato$internet)

dato$m.lebas=as.factor(dato$m.lebas)

data.df=Data

library(ordinal)

library(MASS)

library("MASS")

library(lattice)

library(ggplot2)

library(e1071)

library(caret)

library(rpart)

lib

daramad.class = factor(dato$categuri, ordered = T)

data.df=knn[,37:57]

View(data.df)

attach(dato)

Gender=as.factor(Gender)

```

```

Ostan=as.factor(Ostan)

Faaliat=as.factor(Faaliat)

Madrak=as.factor(Madrak)

oto=as.factor(oto)

internet=as.factor(internet)

m.lebas=as.factor(m.lebas)

Pc=as.factor(Pc)

motor=as.factor(motor)

model=lm(log(daramad)~Gender+Ostan+Faaliat+Madrak+oto+Tedad+T.O+H_Maskan+H_Hamloghl
+H_kalavakhadamat+
H_Tafrihat+H_Hamloghl+H_Pushak+internet+m.lebas+motor+Pc+H_behdasht,data = Data)

summary(model)

library(ordinal)

categurii=as.factor(categuri)

train.index<-sample(c(1:dim(Data)[1]),dim(Data)[1]*0.6)

train.reg<-Data[train.index,]

valid.reg<-Data[-train.index,]

attach(train.reg)

cat=as.factor(train.reg$categuri)

model=polr(cat~Gender+Ostan+Faaliat+Madrak+oto+Tedad+T.O+H_Maskan+H_Hamloghl+H_kalav
akhadamat+

```

```

H_Tafrihat+H_Hamloghl+internet+motor+Pc+H_behdasht,data=      train.reg,na.action      =
na.omit,Hess=TRUE,method = "logistic")

pred <- predict(model,train.reg,type = "class")

confusionMatrix(pred,cat)

valid

attach(valid.reg)

cat=as.factor(valid.reg$categori)

model=polr(cat~Gender+Ostan+Faaliat+Madrak+oto+Tedad+T.O+H_Maskan+H_Hamloghl+H_kalav
akhadamat+
H_Tafrihat+H_Hamloghl+internet+motor+Pc+H_behdasht,data=      valid.reg,na.action      =
na.omit,Hess=TRUE,method = "logistic")

pred <- predict(model,valid.reg,type = "class")

confusionMatrix(pred,cat)

```

## دستورات بخش تحلیل تشخیصی

Discreme

```

library(tidyverse)

library(caret)

theme_set(theme_classic())

data.df[,c(1,3,5,6,7,8,9,10,12:42)] <- lapply(data.df[,c(1,3,5,6,7,8,9,10,12:42)], factor)

train.index<-sample(c(1:dim(data.df)[1]),dim(data.df)[1]*0.6)

```

```
train<-data.df[train.index,]

valid<-data.df[-train.index,]

train.dec=train[,-c(54:59,6,7,13,1,3,5,6,7,8,9,10,12:42)]

valid.dec=valid[,-c(54:59,6,7,13,1,3,5,6,7,8,9,10,12:42)]

preproc.param <- train.dec%>%
  preProcess(method = c("center", "scale"))
```

Transform the data using the estimated parameters

```
train.transformed <- preproc.param %>% predict(train.dec)
```

```
test.transformed <- preproc.param %>% predict(valid.dec)
```

```
library(MASS)
```

```
library(ggplot2)
```

Fit the model

```
model <- lda(categuri~, data = test.transformed)

predictions <- model %>% predict(test.transformed)

as.factor(predict(model)$class)

a=table(predict(model)$class)

predict.dec=factor(predict(model)$class,levels = c(-1.32403275189865, -0.421612018860922,
0.480808714176809, 1.38322944721454 ),labels = c(1,2,3,4))

cat=as.factor(valid$categuri)

confusionMatrix(predict.dec,cat)
```

Model accuracy

```

lda.data <- cbind(train.transformed, predict(model)$x)

class(categuri)

ggplot(lda.data, aes(LD1, LD2)) +
  geom_point(aes(color = categuri))

model <- qda(categuri~, data = train.transformed)

predictions <- model %>% predict(train.transformed)

as.factor(predict(model)$class)

table(predict(model)$class)

predict.dec=factor(predict(model)$class,levels   =  c(-1.34431180433764  ,-0.438419959764337
,0.467471884808962, 1.37336372938226),labels = c(1,2,3,4))

cat=as.factor(train$categuri)

confusionMatrix(predict.dec,cat)

```