Analýza dopadu výšky príjmu na nákup vína, zlata, sladkostí a mäsa

Matej Hrnčiar

Cieľom firmy je maximalizovať zisk. Je pre ňu teda veľmi výhodné odhadnúť svojho zákazníka, aby vedela efektívne cieliť svoju reklamu a míňať čo najmenej financií na propagovanie svojich produktov nesprávnym ľuďom.

Predstavenie datasetu

Našli sme si dataset, ktorý analyzoval rôznych zákazníkov, aby firmám pomohol s tvorbou cielenej reklamy. Dataset obsahuje rôzne atribúty, ktoré opisujú rôznych zákazníkov. Sú v ňom zaznamenaná osobné dáta, ako je vek, príjem, dosiahnuté vzdelanie, rodinný status počet detí v domácnosti. Obsahuje informácie aj odkedy je zákazník zainteresovaný vo firme, kedy si naposledy kúpil nejaký produkt a či vyplnil nejaké sťažnosti v posledných dvoch rokoch. Okrem toho obsahuje údaje o sume minutých na produkty ako sú víno, ovocie, ryby, mäso, sladkosti a zlato. Tiež sa v ňom nachádzajú dáta o miestach nákupu a využívaní zliav na kúpu produktov. Máme teda k dispozícii dostatok informácií o zákazníkovi, ktoré o ňom môžeme skúmať. Môžeme si vymýšľať rôzne hypotézy a prichádzať na spôsoby ako spolu jednotlivé stĺpce súvisia, ale vybrali sme si jednoducho overiť, že či ľudia s vyšším príjmom míňajú viac na produkty.

Pozrieme sa však na niektoré produkty, konkrétne víno, zlato, sladkosti a mäso. Naša prvotná teória je taká, že ľudia s vyšším príjmom budú míňať viac na zlato, keďže im zostane viac prostriedkov na investovanie. Pri sladkostiach a mäse si myslíme, že rozdiel nebude veľký, keďže sú to celkom bežné produkty a sú každému dostupné. Tiež rovnakú situáciu predpokladáme aj pri víne. Samotná hypotéza nám zmysel dáva, veď keď niekto viac zarobí, môže predsa viac minúť. A ak niekto viac zarobí, tak mu zostane viac peňazí, ktoré môže využiť na neesenciálne produkty. Čiže nejaká korelácia v našom tvrdení je. Nemyslíme si však, že by išlo o kauzalitu. Aj menej zarábajúci ľudia sa môžu raz rozhodnúť investovať do zlata a možno menej jesť v daný mesiac. Alebo nastane špeciálna príležitosť, pri ktorej budú potrebovať viac vína a aj občerstvenia. Na overenie nám stačia konkrétne stĺpce, ktoré náš dataset obsahuje.

Dataset sme si načítali do prostredia a prezreli sme si ho. Obsahuje 29 stĺpcov a 2240 riadkov. Tri stĺpce majú textový typ a všetky ostatné stĺpce sú numerické. Všetky stĺpce k analýze máme k dispozícii, netreba teda robiť žiadne transformácie. Pre našu analýzu nás nebudú zaujímať všetky stĺpce. Budú nás zaujímať len stĺpce *Income, Kidhome, Teenhome, MntWines, MntMeatProducts, MntSweetProducts* a *MntGoldProds*.

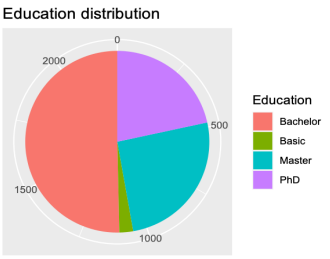
Základná analýza premenných

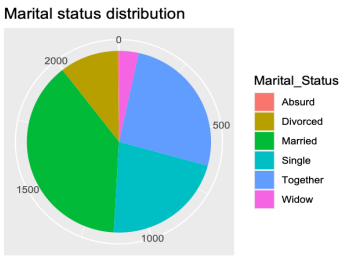
Keď si zobrazíme nejaké základné štatistiky o stĺpcoch, vidíme, že stĺpce našich produktov nemajú na prvý pohľad problémy. Neobsahujú žiadne chýbajúce hodnoty a aj vypočítané štatistické hodnoty vyzerajú celkom správne, teda nevidíme v nich nejaké divné a veľké hodnoty. Pri stĺpci *Income* je však už tá situácia iná. Vidíme, že stĺpec obsahuje 24

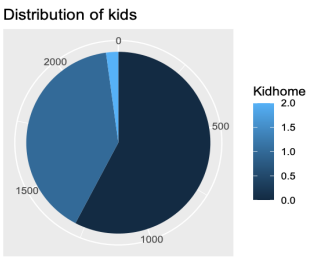
chýbajúcich hodnôt. Keď sa na tieto hodnoty bližšie pozrieme, vidíme, že aj keď nemajú žiadnu hodnotu pri príjme, minuli peniaze na produkty. Najlepším riešením nám teda prišlo dosadiť priemer stĺpca *Income* namiesto NA hodnôt. Okrem toho má stĺpec *Income* nejaké vychýlené hodnoty. Tieto hodnoty musíme predom odstrániť, aby sme predišli nepresným výsledkom pri lineárnej regresii. Urobíme to pomocou *interquality range (IQR)*. Vypočítame si kvantil stĺpca a všetky merania pod hodnotou 25. kvartilu - 1.5*IQR a nad hodnotou 75. kvartilu + 1.5*IQR odstránime. Takto sa zbavíme vychýlených hodnôt, ktorých bolo v stĺpci 8.

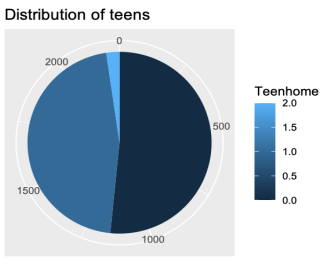
Spravíme si základné deskriptívne štatistiky premenných pomocou koláčových a kernel density estimation grafov. Môžeme vidieť, že najväčšie zastúpenie v datasete majú osoby bez detí a bez tínedžerov. Taktiež môžeme sledovať, že ľudia bez detí majú v priemere vyšší príjem ako ľudia s jedným alebo dvomi deťmi. Príjem pri tínedžeroch je rozdelený na dva vrcholy: ľudia, ktorí nemajú v domácnosti tínedžerov a majú v priemere nižšie platy, pravdepodobne lebo si nemôžu dovoliť živiť ďalšieho člena rodiny, a potom sú ľudia bez tínedžerov, ktorí majú v priemere vyššie platy.

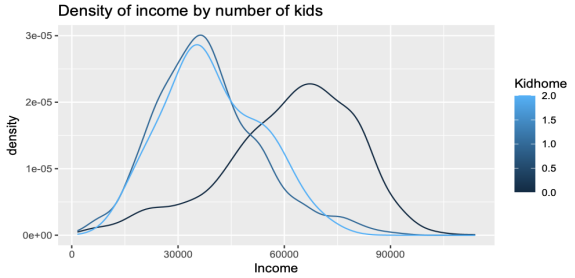
Ďalšie skúmané premenné máme najvyššiu dosiahnutú úroveň vzdelania a rodinný stav. Väčšinu vzorky tvoria ľudia s titulom bakalára, ale je pozoruhodné, že dataset tvorí aj pomerne veľké množstvo ľudí s titulom PhD. Rodinný stav zobrazuje, že viac ako tretina vzorky je ženatá, následne sú ľudia, ktorí žijú spolu ale nie sú manželia, slobodní a najmenšie zastúpenia majú rozvedení, ovdovení alebo ľudia, ktorí uviedli absurdný rodinný stav. Tieto dve premenné by mohli byť zajímavé na preskúmanie v korelácii s príjmom a výdavkami na rôzne druhy produktov, ale to už zasahuje mimo rozsah našej práce.

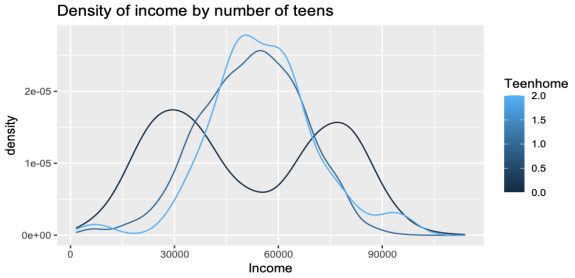












Hypotéza

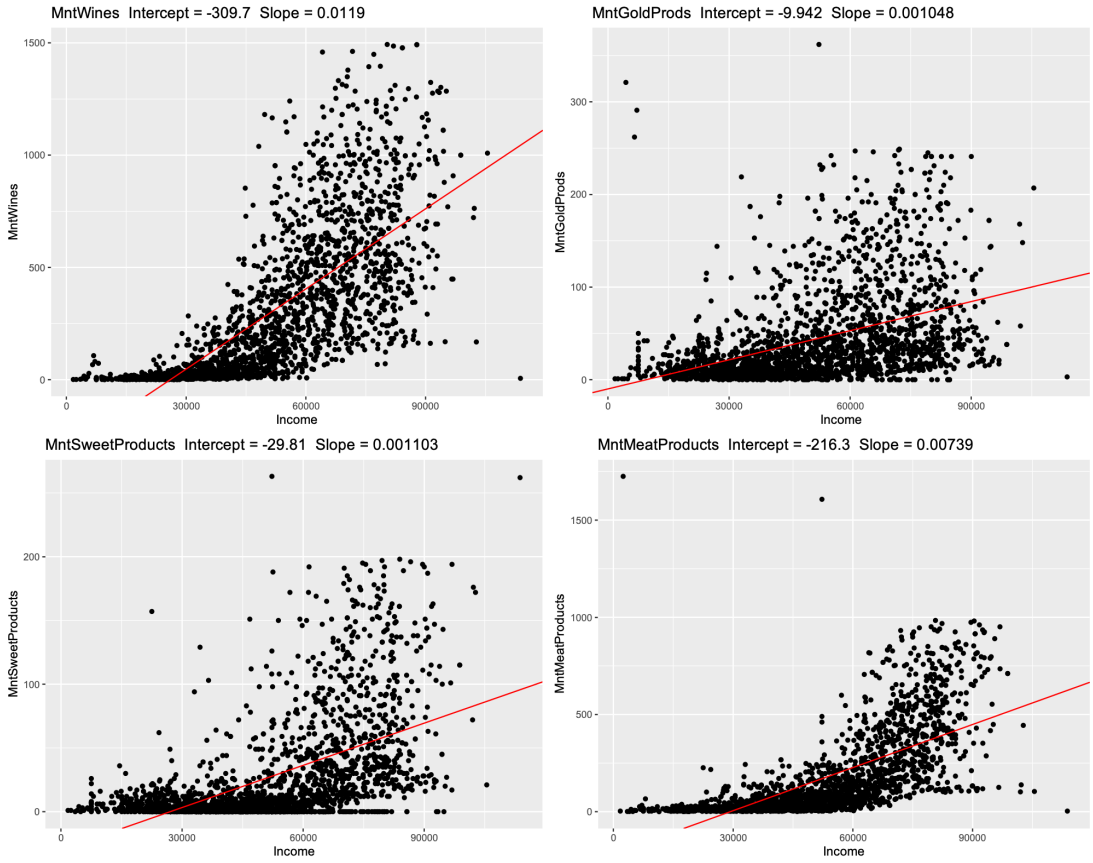
Hodnoty teda máme upravené a môžeme sa pozrieť na lineárnu regresiu. Pre našu analýzu sme si stanovili hypotézy:

H0: ľudia s vyšším príjmom míňajú v priemere rovnako veľa peňazí na produkty ako víno, mäso, sladkosti a zlato ako ľudia s nižším príjmom

H1: ľudia s vyšším príjmom míňajú v priemere viac peňazí na produkty ako víno, mäso, sladkosti a zlato ako ľudia s nižším príjmom

Najprv sa pozrieme na základné korelácie medzi príjmami a výdavkami na produkty bez prihliadania na počet detí a tínedžerov v domácnosti. Pre overenie si vygenerujeme 40 podmnožín, ktoré budú reprezentovať náhodných 50% meraní. Čiže najprv si náhodne vyberieme polovicu meraní z datasetu. Tieto hodnoty zoradíme od, aby sme iterovali od najnižšej hodnoty po najvyššiu a zapíšeme do listu. Toto urobíme 40 krát a tak nám vznikne našich 40 vzoriek. Keď už máme naše vzorky, môžeme vytvoriť model. Tu pre každú vzorku len fitneme lineárny model k dátam.

V ďalšom kroku si vyextraktujeme β koeficienty. Následne vypočítame sklon a intercept a nakoniec ešte vypočítame residual sum of squares a residual mean error pre každú množinu, aby sme zistili celkovú chybovosť modelu. Tá má pri každom produkte malý rozptyl, hodnoty koeficientov nám lietajú čo najmenej a teda model máme dobre namapovaný. Nakoniec ešte overíme stabilitu modelu. Urobíme si t-test pre každý model, ktorý nám overí, či je medzi príjmom a daným produktom nejaký vzťah. Pozrieme teda, že aká veľká je odchýlka od priamky grafu. Hodnota t testu nám pre každý produkt vychádza veľmi blízko k



nule, pri víne to býva 0.1, pri zlate a sladkostiach 0.01 a pri mäse 0.07. Vykonanú regresiu si môžeme dať do grafov, z čoho nám vznikli nasledujúce výsledky:

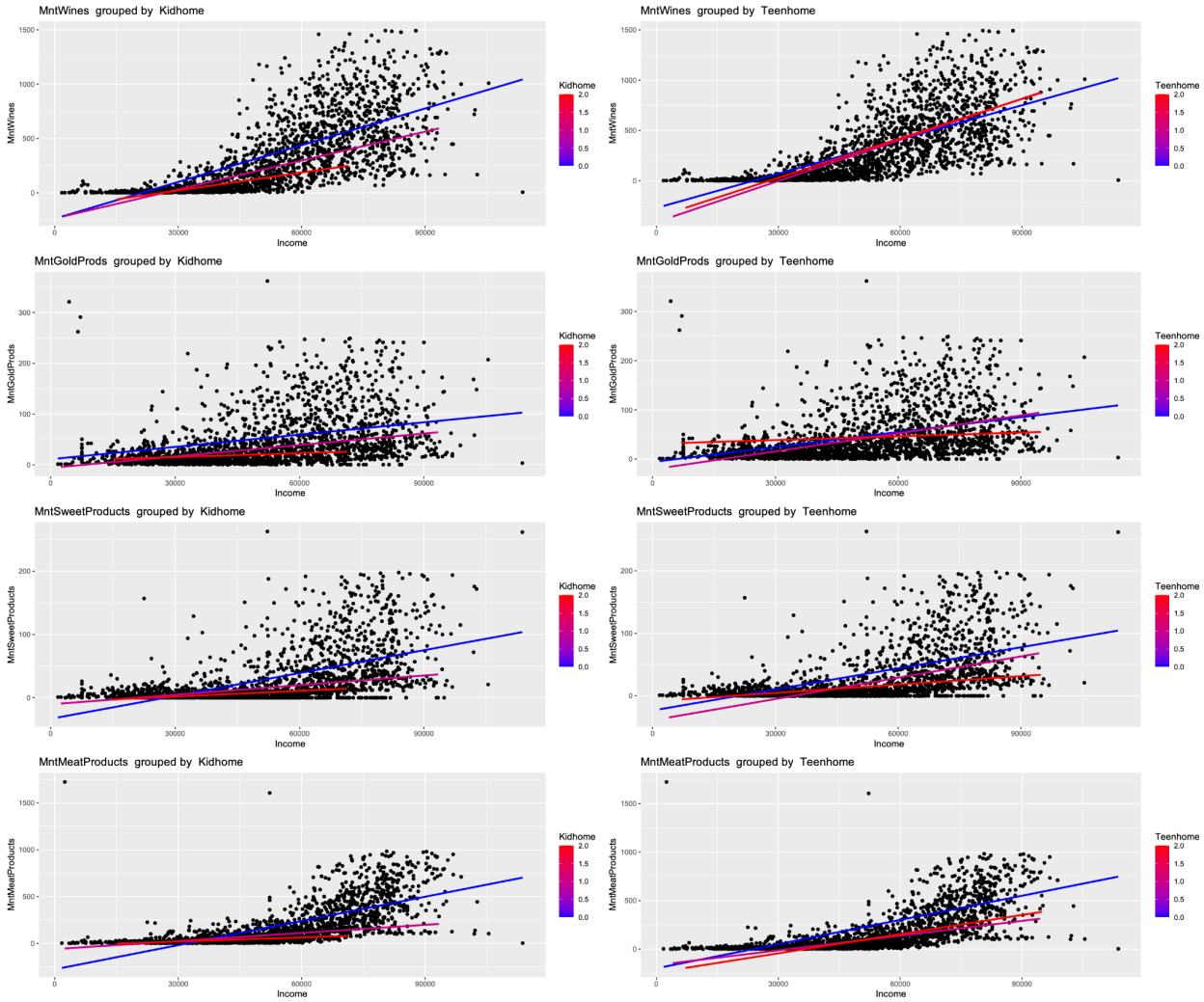
Pozorovania

Vidíme, že v modeli sa mohli nachádzať nejaké vychýlené hodnoty. Najlepšie to je viditeľné na mäsových alebo aj zlatých produktoch, kedy ľudia so skoro nulovým príjmom míňali veľa peňazí na dané produkty. Týchto situácií však nie je veľa a teda stabilitu modelu veľmi nenarušili. Je teda vidno, že naša hypotéza sa nám potvrdila - ľudia s vyšším príjmom míňajú v priemere viac na peňazí na všetky vybrané produkty, a aj keď sme takýto výsledok očakávali, našli sme pár vecí, ktoré nás aj prekvapili.

Na víno ľudia míňajú veľa peňazí a je oveľa obľúbenejšie ako mäso alebo sladkosti, čo sa dalo očakávať, ale s rastúcim príjmom rastie nákup vína oveľa rýchlejšie ako pri všetkých ostatných produktoch. Predpokladali sme, že ľudia s vyšším príjmom budú nakupovať veľa produktov zo zlata, ale táto krivka rastie najpomalšie spomedzi všetkých produktov, čo nám napovedá, že zlato, minimálne pre skúmanú vzorku, nie je až tak populárne.

Rozšírenie

Zaujímavým rozšírením môže byť skúmanie nákupu produktov rozdelené podľa počtu detí a tínedžerov v domácnosti:



Tu môžeme vidieť, že ľudia bez detí míňajú v priemere viac na všetky skúmané druhy produktov, avšak pri tínedžeroch sú výsledky rôzne. Nákup sladkostí a mäsových výrobkov nie sú veľmi odlišné od detí, ale ľudia s jedným tínedžerom v domácnosti majú tendenciu míňať približne rovnaké množstvo peňazí na zlato ako ľudia bez tínedžerov, a pri víne dokonca míňajú viac ak majú aspoň jedného tínedžera, aj keď nie o veľa. Tento výsledok však nemusí byť správny, nakoľko pri základnej analýze premenných sme mohli pozorovať, že skupiny s dvoma deťmi alebo tínedžermi sú veľmi malé a teda nie sú veľmi spoľahlivé pre regresiu.

Ďalšími rozšíreniami môže byť skúmanie nákupu produktov rozdelené podľa najvyššej dosiahnutej úrovne vzdelania alebo rodinného stavu, ktoré môžu taktiež odhaliť zaujímavé trendy, avšak skúmanie týchto premenných je nad rámec tejto práce. Pôvodnú hypotézu sa nám podarilo dokázať pomocou regresných modelov a grafov a predpokladáme, že aj pri delení podľa spomínaných premenných by bol trend rastúci, aj keď by nám mohol ukázať, ktoré skupiny ľudí míňajú viac peňazí na nejaký druh produktu.