

به نام خدا



تمرین فصل دوم داده کاوی

استاد مربوطه: دکتر غلامی

پرستو غلامی

۹۹۳۶۱۳۰۴۸

اسفند ۱۴۰۲

با توجه به اینکه قرار است همبستگی میان تمامی ستون‌ها به صورت کامل و دقیق سنجیده شود، می‌توان این کار را با کمک یک حلقه for به سادگی انجام داد.

بعد از انجام این کار به همبستگی زیر میان ستون‌های مختلف می‌پردازیم.

	Name	Insulation	Temperature	Heating_Oil	Num_Occupants	Avg_Age	Home_Size
0	Insulation	1.0	-0.793696	0.736097	-0.012567	0.642982	0.200712
1	Temperature	-0.793696	1.0	-0.77366	0.012519	-0.672579	-0.213939
2	Heating_Oil	0.736097	-0.77366	1.0	-0.041635	0.847891	0.381191
3	Num_Occupants	-0.012567	0.012519	-0.041635	1.0	-0.048034	-0.022534
4	Avg_Age	0.642982	-0.672579	0.847891	-0.048034	1.0	0.306557
5	Home_Size	0.200712	-0.213939	0.381191	-0.022534	0.306557	1.0

۱. بررسی همبستگی‌های ستون Insulation:

a. **Temperature**: طبق همبستگی موجود، این دو مورد باهم همبستگی منفی دارند. یعنی اگر میزان عایق بیشتر باشد، دما بیرون کمتر و اگر میزان عایق کمتر باشد، دما بیرون بیشتر است. از آنجایی که هدف حفظ دمای خانه است، و با سرد شدن هوا میزان عایق بیشتر می‌شود، این همبستگی نسبتاً بالا، منطقی است. از آنجایی هم که اگر هوای خارج خانه گرم باشد، نیازی به حفظ حرارت داخل خانه نیست، نیازی به حفظ این دما نبوده و از عایق کمتری استفاده شده است. البته اگر بخواهیم به دید منطقی‌تر به این قضیه نگاه کنیم، برای حفظ دمای پایین (سرما) نیز نیاز به عایق داریم که احتمالاً دلیل اینکه این همبستگی از این عدد بیشتر نشده است همین موضوع باشد.

b. **Heating_Oil**: با توجه به ارتباط همبستگی‌های بین این دو، هرچه مقدار خرید سوخت بیشتر باشد، عایق خانه نیز بیشتر است. همچنین اگر خرید سوخت کم شود، عایق نیز کاهش می‌یابد. از آنجایی که خرید سوخت، برای افزایش دمای خانه است، اگر خانه به خوبی عایق شده باشد، سوخت کمتری نیاز به مصرف است. همچنین اگر خانه به خوبی عایق نباشد، نیاز است که سوخت بیشتری استفاده شود که این موضوع از نظر مفهومی باید همبستگی منفی داشته باشد. بنظر می‌رسد همبستگی مستقیم میان این دو، به دلیل همبستگی بالای منفی هردو این موارد با دمای خارج از خانه یعنی پارامتر **Temperature** است. یعنی این دو باید

از نظر مفهومی باهم همبستگی معکوس داشته باشند اما به دلیل اینکه هردو با Temperature یک نوع همبستگی دارند، باعث شده این دو نیز همبستگی داشته باشند.

c. Num_Occupants: این دو پارامتر باهم هیچ همبستگی ندارند. همچنین از نظر مفهومی نیز نباید باهم همبستگی داشته باشند و اگر همبستگی داشتند این موضوع جای شک داشت. این موضوع یعنی افزایش یا کاهش تعداد اعضای خانوار، باعث افزایش یا کاهش میزان عایق خانه نمی‌شود!

d. Avg_Age: این همبستگی مثبت بیان می‌دارد که اگر سن افراد یک خانه افزایش پیدا کند، میزان عایق نیز بیشتر می‌شود. این موضوع شدیداً عجیب است. این دو نیز از نظر مفهومی نباید با یکدیگر همبستگی داشته باشند. اما از آنجایی که افرادی که سن بالاتر دارند ممکن است در صرفه جویی حساس‌تر باشند (به علت کاهش درآمد و بازنشستگی) برای خرج نکردن در خرید سوخت ممکن است از عایق استفاده کنند. اما باز هم این همبستگی بالا منطقی نیست.

e. Home_Size: این دو باهم همبستگی مستقیم اما بسیار کم دارند. این همبستگی اشاره دارد تا حدی اگر سائز خانه بیشتر شود، عایق نیز بیشتر می‌شود. اما مقدار این همبستگی بسیار اندک است. از آنجایی که در خانه‌های بزرگ حرارت راحت خارج می‌شود، منطقی‌تر است که عایق بیشتری استفاده شود. اما چون درآمد بالایی دارند، ممکن است به حفظ دما با عایق بی توجهی کرده اما با استفاده از سوزاندن سوخت، دمای خانه را افزایش دهند.

۲. بررسی همبستگی‌های ستون Temperature:

a. Heating_Oil: این دو بیان می‌دارد که هرچی دما کمتر باشد، مصرف سوخت بیشتر و هرچی دما بیشتر باشد مصرف سوخت کمتر است. این باهم ارتباط معکوس دارند. منطقی نیز هست زیرا با کاهش دما نیاز است سوخت بیشتری مصرف شود.

b. Num_Occupants: از آنجایی که افزایش دمای خارج از خانه با تعداد اعضای یک خانواده از نظر منطقی هیچ رابطه‌ای ندارد، در این داده‌ها نیز این دو باهم هیچ وابستگی ندارند. یعنی افزایش یا کاهش تعداد اعضای خانواده با دمای بیرون هیچ ارتباطی ندارد و کاملاً بی ربط کم و زیاد می‌شود (:

c. Avg_Age: این دو در این داده‌ها باهم ارتباط معکوس دارند. یعنی در داده‌ها با افزایش سن افراد یک خانه، دمای خانه کاهش می‌یابد. و با افزایش دما میانگین سن کاهش می‌یابد. اما این موضوع از نظر منطقی به هیچ عنوان قابل توجیه نیست و به دلایل دیگر ممکن است این همبستگی به وجود آمده باشد. به همین دلیل میزان همبستگی در این مورد به هیچ عنوان قابل اتکا نیست.

d. Home_Size: این دو باهم ارتباط معکوس بسیار اندکی دارند. اما به صورت منطقی هیچ ارتباطی باهم ندارند. همچنین این ارتباط بسیار اندک است و افزایش و کاهش سائز خانه با دمای خارج خانه باهم ارتباطی ندارند.

۳. بررسی همبستگی‌های ستون Heating_Oil:

a. Num_Occupants: همبستگی میان این دو بسیار کم است. یعنی افزایش یا کاهش افراد داخل خانه در میزان خرید سوخت اثری ندارد. البته از آنجایی که اگر افراد بیشتری در یک خانه زندگی کنند، دمای آن خانه بیشتر حفظ می‌شود، توقع می‌رود که مصرف سوخت باید کمتر باشد. توقع بر این بود که این دو بیشتر باهم همبستگی داشته باشند اما در این مورد هیچ همبستگی باهم ندارند که مقداری عجیب است.

b. Avg_Age: این دو همبستگی مستقیم و بالایی دارند. یعنی با افزایش میانگین سن میزان خرید سوخت بیشتر می‌شود. همچنین با کاهش میانگین سن، میزان خرید سوخت کمتر می‌شود. این دو همبستگی بسیار بالایی دارند و این میزان، بیشترین همبستگی میان تمامی پارامترهای موجود است. اما از نظر منطقی دلیل محکمی برای آن وجود ندارد. صرفاً ممکن است افراد مسن‌تر به دلیل اینکه بیشتر احساس سرما می‌کنند بیشتر از سوخت استفاده کنند اما این ارتباط به قدری قوی نیست که باعث شود چنین همبستگی بالایی ایجاد شود.

c. Home_Size: این دو بیشتری میزان همبستگی را در پارامتر مورد بررسی Home_Size دارند. این موضوع تا حدی منطقی است. زیرا هرچه یک خانه بزرگ‌تر باشد، سوخت بیشتری برای گرم کردن آن مورد نیاز است. عدد همبستگی این دو تقریباً مستقیم است اما توقع می‌رفت که این عدد با توجه به ارتباط منطقی محکم بین این دو بیشتر نیز باشد. در این مورد اگر سائز خانه زیاد شود، مصرف سوخت بیشتر و اگر سائز خانه کم شود، مصرف سوخت کم می‌شود.

۴. بررسی همبستگی‌های ستون Num_Occupants:

a. Avg_Age: این دو باهم هیچ ارتباطی ندارند. یعنی اگر میانگین سن کم یا زیاد شود، لزومی ندارد که تعداد اعضای یک خانوار نیز کم یا زیاد شود. البته که رابطه منطقی زیادی بین این دو نیز وجود ندارد. زیرا نمی‌توان گفت فرزندان و والدین چه سنی دارند.

b. Home_Size: این دو باهم هیچ گونه همبستگی ندارند. یعنی کم یا زیاد شدن جمعیت یک خانواده بر کوچک یا بزرگ شدن سائز خانه اثری ندارد. البته از نظر منطقی، باید یک ارتباطی میان این دو وجود داشته باشد. یعنی خانواده پر جمعیت‌تر، خانه بزرگ‌تر و خانواده کم جمعیت‌تر، خان کوچک‌تر داشته باشند. نکته‌ای که مهم است این عدم همبستگی ممکن است

به چند دلیل باشد. دلیل نخست اینکه، نوع دسته بندی داده اندازه خانه صحیح نباشد. این موضوع می تواند اثر شدیدی بر همبستگی بزارد. دلیل دوم اینکه می تواند به دلیلی مانند پولدار بودن یا نبودن، افراد را مجبور به خرید خانه های دیگر کند. مثلا افراد کم جمعیت که پولدار هستند، خانه های بزرگ و افراد پر جمعیت که فقیر هستند، خانه کوچک داشته باشند. این مورد دقیقا اثر داده سومی است که در داده های ما موجود نیست و نمی توانیم درباره آن ها نظر دهیم.

۵. بررسی همبستگی های ستون Avg_Age:

a. Home_Size: این دو همبستگی مستقیم و کمی باهم دارند. یعنی ممکن است با افزایش میانگین سنی، سایز خانه بزرگ تر شود. این دو از نظر منطقی اندکی باهم ارتباط دارند، زیرا تا حدی با افزایش سن، افراد می توانند خانه های بزرگ تری تهیه کنند. اما بعد از آن به دلیل بازنشستگی، دیگر امکان تهیه خانه بزرگ تر را ندارند. بنابراین اینکه تا حدی این عدد همبستگی نشان دهد، منطقی است.