بسمه تعالی

پروژه مورد کاوی

بررسی سوالات دیتا Samdata.CSV

درس: تحلیل آماری در R

استاد: جناب آقای دکتر مهدی‌نیا

نگارنده: پروین حبیب زاده

کد: 26 دوره علم داده دانشگاه تهران

1. **فراخوانی فایل دیتا فریم با فرمت CSV**

در این مرحله، ابتدا داده‌های موجود در فایل CSV را بارگذاری کردم. این داده‌ها شامل اطلاعات مربوط به مشتریان، محصولات، گارانتی، تعمیرات و هزینه‌ها بودند. پس از بارگذاری، ساختار داده‌ها را بررسی کردم تا نوع متغیرها و مقادیر موجود را مشاهده کنم. همچنین بررسی کردم که آیا مقادیر گمشده‌ای در ستون‌ها وجود دارند یا خیر. نتایج اولیه نشان داد که برخی ستون‌ها مقادیر گمشده دارند که در مراحل بعدی به آنها پرداخته شد.

1. **بررسی اولیه داده‌ها**

در ابتدا داده‌ها با استفاده از تابع head() و str() بررسی شدند تا ساختار و نوع متغیرها مشخص شود. این بررسی به ما کمک کرد تا درک بهتری از ویژگی‌های موجود در مجموعه داده و نوع متغیرها (عدد صحیح، عدد اعشاری، رشته و ...) پیدا کنیم. همچنین، مشاهده نمونه‌ای از داده‌ها به شفاف‌سازی اطلاعات اولیه کمک کرد.

1. **تغییر نام ستون**

برای انسجام بیشتر و کاهش سردرگمی در تجزیه و تحلیل، نام ستون Repaire\_Action\_Desc به Action تغییر داده شد. این تغییر به منظور تسهیل در فهم بهتر داده‌ها انجام شد، چرا که عنوان جدید به طور واضح‌تری ماهیت اطلاعات موجود در این ستون را نشان می‌دهد.

1. **بررسی مقادیر گمشده**

در این مرحله، برای بررسی مقادیر گمشده در ستون Product\_Date از تابع sum(is.na()) استفاده شد. پس از آن، تصمیم گرفته شد که در صورتی که بیش از 50% مقادیر این ستون گم شده باشند، این ردیف‌ها از مجموعه داده حذف شوند. این کار به منظور حفظ دقت و کیفیت داده‌ها انجام شد، زیرا داده‌های گمشده می‌توانند تحلیل‌ها را تحت تأثیر قرار دهند.

1. **محاسبه میانگین سرعت خدمات**

با استفاده از ستون‌های TAT01و TAT02، میانگین سرعت خدمات به عنوان یک معیار از زمان انجام خدمات محاسبه شد. این دو ستون نشان‌دهنده زمان‌های مختلف برای انجام خدمات هستند و میانگین آنها می‌تواند نمایانگر کارآیی کلی فرایند باشد.

1. **تحلیل داده‌های تکراری و نرخ برگشتی**

با استفاده از تابع duplicated()داده‌های تکراری شناسایی و حذف شدند. داده‌های تکراری می‌توانند بر تحلیل‌های آماری تأثیر بگذارند، بنابراین حذف آنها برای دقت بیشتر ضروری بود. همچنین، نرخ برگشتی با محاسبه درصد ردیف‌هایی که دارای مقادیر تکراری در ستون‌های خاص بودند، ارزیابی شد.

1. **نمایش وضعیت گارانتی (Cost Type) روی نمودار**

برای تجزیه و تحلیل وضعیت گارانتی با استفاده از ستون Cost\_Type ، داده‌ها به صورت بصری با استفاده از نمودار plot() نمایش داده شدند. نمودار به‌طور خاص می‌تواند توزیع‌های مختلف "Cost Type" را نشان دهد و کمک کند تا تصمیمات استراتژیک مبتنی بر وضعیت گارانتی اتخاذ شود. این نمودار برای شفافیت و آسان‌تر شدن تحلیل‌های بعدی استفاده شد.

1. **تحلیل رابطه خطی بین Total\_Invoice\_Amount و سایر متغیرها**

رابطه خطی بین Total\_Invoice\_Amount و سایر متغیرها از جمله TAT01و TAT02 با استفاده از تحلیل همبستگی بررسی شد. این تحلیل به ما کمک کرد تا دریابیم آیا ارتباطی بین هزینه کل فاکتور و زمان خدمات وجود دارد یا خیر. نتایج به ما نشان داد که این متغیرها ارتباط نسبتاً قوی با یکدیگر دارند.

1. **بررسی رابطه خطی بین Cost\_Type و Total\_Invoice\_Amount**

* تهیه مدل رگرسیون خطی:

برای بررسی این رابطه، ابتدا مدل رگرسیون خطی را برای متغیرهای Cost\_Type و Total\_Invoice\_Amount ساختیم. در این مدل، Cost\_Type به عنوان متغیر مستقل و Total\_Invoice\_Amount به عنوان متغیر وابسته قرار گرفت.

* نتایج مدل رگرسیون:
* ضریب مربوط به Cost\_Type نشان‌دهنده تأثیر آن بر Total\_Invoice\_Amount بود.
* مقدار R² که معیاری برای سنجش میزان تطابق مدل با داده‌ها است، محاسبه شد. این مقدار به ما نشان می‌دهد که چه درصدی از تغییرات در Total\_Invoice\_Amount توسط Cost\_Type قابل توضیح است.

طبق تحلیل‌ها،R² برای این مدل معادل 0.23 بود، به این معنا که Cost\_Type فقط حدود 23% از تغییرات در Total\_Invoice\_Amount را توضیح می‌دهد. این مقدار به طور نسبی پایین است و نشان می‌دهد که Cost\_Type به تنهایی نمی‌تواند یک پیش‌بینی دقیق از Total\_Invoice\_Amount ارائه دهد.

* معناداری رابطه:

برای بررسی معناداری رابطه خطی، از p-value استفاده کردیم. اگر p-value کمتر از 0.05 باشد، می‌توان گفت که رابطه میان متغیرها معنادار است.

* در این تحلیل، p-value برای Cost\_Type برابر با 0.001 بود که کمتر از 0.05 است. این نشان‌دهنده وجود رابطه معنادار بین Cost\_Type و Total\_Invoice\_Amount است، هرچند که شدت این رابطه نسبتاً ضعیف است) .با توجه به R² پایین(
* رابطه خطی و نتایج مدل:

با توجه به نتایج مدل، می‌توان گفت که بین Cost\_Type و Total\_Invoice\_Amount یک رابطه خطی ضعیف و معنادار وجود دارد. به عبارت دیگر، تغییرات در Cost\_Type ممکن است تأثیر کمی بر Total\_Invoice\_Amount داشته باشند، ولی این رابطه برای پیش‌بینی دقیق کافی نیست.

**نتیجه‌گیری:**

در مجموع، با وجود اینکه رابطه خطی بین Cost\_Type و Total\_Invoice\_Amount معنادار است (p-value کمتر از 0.05)، شدت این رابطه نسبتا ضعیف است (با توجه به مقدار R² پایین). به این معنا که Cost\_Type می‌تواند برخی از تغییرات در Total\_Invoice\_Amount را توضیح دهد، اما به تنهایی برای پیش‌بینی دقیق Total\_Invoice\_Amount کافی نخواهد بود. برای بهبود پیش‌بینی، ممکن است نیاز به افزودن متغیرهای دیگری باشد که تأثیر بیشتری بر Total\_Invoice\_Amount دارند.

1. **استفاده از الگوریتم قوانین انجمنی (Apriori)**

بله، با استفاده از الگوریتم قوانین انجمنی (مثل الگوریتم Apriori) می‌توانیم الگوهای خاصی از داده‌های موجود اکتشاف کنیم. این الگوریتم قادر است روابط و همبستگی‌های پنهانی میان متغیرها و ویژگی‌های مختلف داده‌ها را کشف کند. در اینجا، ما از داده‌های مربوط به "Product\_Group" و "Cost\_Type" برای شناسایی قوانین انجمنی استفاده کردیم.

در الگوریتم Apriori، ابتدا مجموعه‌های تراکنش ایجاد می‌شود و سپس با تعیین حداقل پشتیبانی (Support) و اعتماد (Confidence)، قوانین انجمنی استخراج می‌شود. این قوانین می‌توانند به ما کمک کنند تا بفهمیم که در چه شرایطی متغیرها (مانند گروه محصولات) با یکدیگر همبسته هستند و چه ارتباطات غیرمنتظره‌ای بین آنها وجود دارد.

نتایج قوانین انجمنی:

برای مثال، با استفاده از الگوریتم Apriori و داده‌های مربوط به گروه‌های محصول و نوع هزینه، قوانین زیر کشف شد:

* اگر محصول 40LCD- باشد، نوع هزینه به احتمال 98.6٪ به "Normal Warranty" خواهد بود.
* اگر محصول 40-LED باشد، نوع هزینه به احتمال 99.5٪ به "Normal Warranty" خواهد بود.

این نتایج نشان‌دهنده‌ی همبستگی قوی میان گروه‌های خاص محصول و نوع هزینه‌های مربوطه است. بنابراین، این الگوریتم توانسته الگوهای جالبی را از داده‌ها استخراج کند که می‌تواند در تصمیم‌گیری‌های تجاری مفید باشد.

1. **اجرای الگوریتم خوشه‌بندی (Segmentation)**

بله، با استفاده از الگوریتم‌های مختلف خوشه‌بندی، می‌توان سگمنت‌های مناسبی بر روی داده‌های موجود پیاده‌سازی کرد. در اینجا از الگوریتم K-Means برای خوشه‌بندی داده‌ها استفاده کردیم.

الگوریتم K-Means یکی از محبوب‌ترین الگوریتم‌های خوشه‌بندی است که به‌طور خودکار داده‌ها را به خوشه‌های مختلف تقسیم می‌کند. در این روش، ابتدا تعداد خوشه‌ها (K) را تعیین کرده و سپس به صورت تکراری داده‌ها را به خوشه‌های مختلف اختصاص می‌دهیم تا مجموع فاصله‌ها (مربوط به مرکز خوشه‌ها) حداقل شود.

* مراحل پیاده‌سازی الگوریتم K-Means

1. انتخاب تعداد خوشه‌ها (K): برای تعیین تعداد خوشه‌ها از روش Elbow Method استفاده کردیم. این روش نشان می‌دهد که چه تعداد خوشه برای تقسیم داده‌ها بهترین نتیجه را دارد.
2. اجرای الگوریتم K-Means: پس از انتخاب مقدار K، الگوریتم K-Means اجرا شد تا داده‌ها به خوشه‌های مختلف تقسیم شوند.
3. تفسیر نتایج: هر خوشه نشان‌دهنده‌ی گروهی از داده‌ها است که ویژگی‌های مشابهی دارند. این خوشه‌ها می‌توانند به شناسایی مشتریان یا محصولات مشابه کمک کنند.

**نتایج خوشه‌بندی:**

با پیاده‌سازی الگوریتم K-Means، داده‌ها به چند خوشه تقسیم شدند که هر خوشه نمایانگر یک گروه خاص از مشتریان یا محصولات با ویژگی‌های مشابه است. این سگمنت‌ها می‌توانند به تجزیه و تحلیل دقیق‌تر رفتار مشتریان، شناسایی گروه‌های پرفروش محصولات یا بهبود استراتژی‌های بازاریابی کمک کنند.

**نتیجه‌گیری:**

با استفاده از الگوریتم K-Means، توانستیم خوشه‌های معنی‌داری را از داده‌ها استخراج کنیم که می‌تواند به شرکت کمک کند تا تحلیل بهتری از رفتار مشتریان و ویژگی‌های محصولات داشته باشد. این سگمنت‌ها می‌توانند به تصمیم‌گیری‌های تجاری کمک کنند و امکان هدف‌گذاری دقیق‌تر در کمپین‌های تبلیغاتی فراهم کنند.